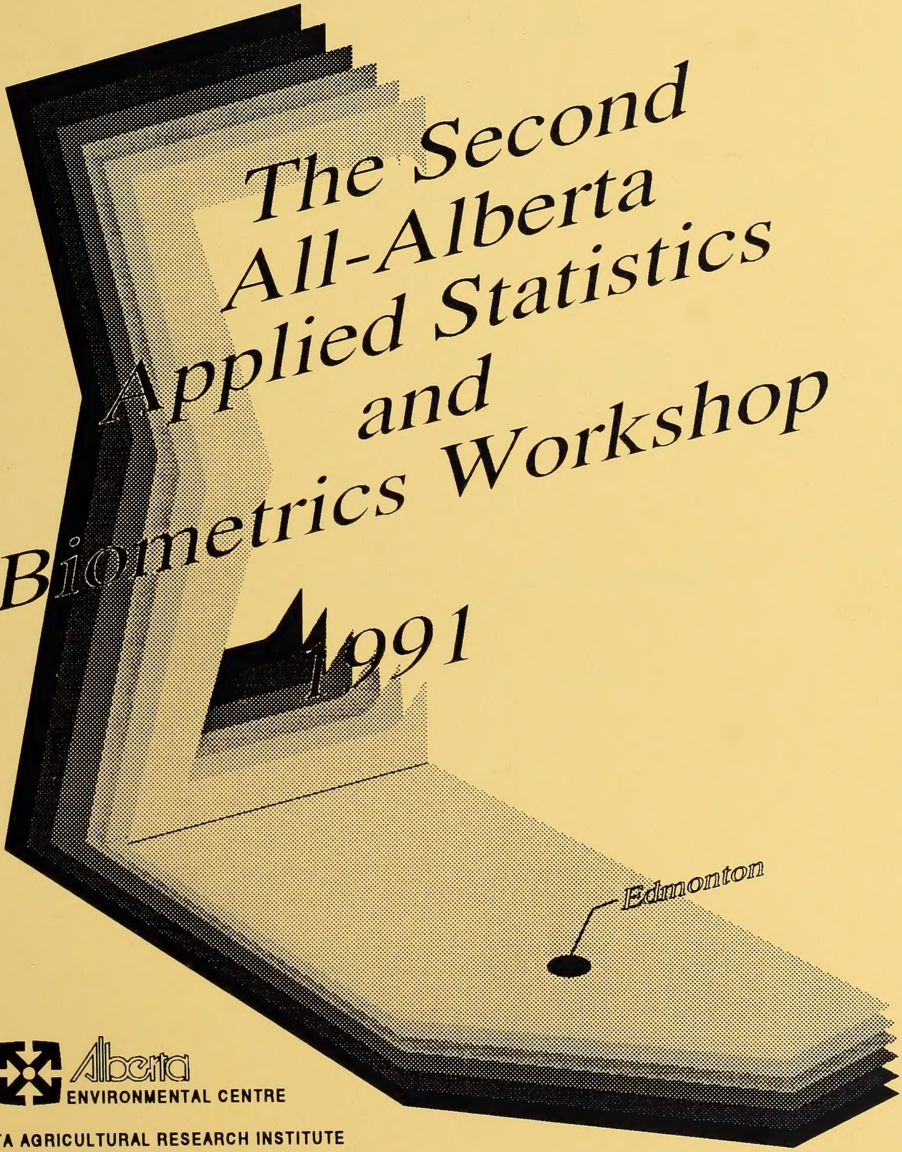
AECV-92P1

Workshop Program & Proceedings

The Second All-Alberta Applied Statistics and Biometrics Workshop 1991



Alberta
ENVIRONMENTAL CENTRE

ALBERTA AGRICULTURAL RESEARCH INSTITUTE

Alberta
AGRICULTURE

Program and Proceedings of the Second All-Alberta Applied Statistics and Biometrics Workshop

Organized by

L.Z. Florence and L.A. Goonewardene

October 21-22, 1991

Crowne Plaza

10111 Bellamy Hill

Edmonton, Alberta

Canada

Alberta ENVIRONMENTAL CENTRE

ALBERTA AGRICULTURAL RESEARCH INSTITUTE

Alberta AGRICULTURE

DISCLAIMER

The contents of these Proceedings reflect only the opinions of the individual authors and must not be construed in any manner to represent the opinions or policies, official or unofficial, of Alberta Agriculture, the Alberta Environmental Centre or Alberta Environment.

COPYRIGHT STATEMENT

Copies of the complete proceedings may be obtained by contacting:

Alberta Environmental Centre
Postal Bag 4000
Vegreville, Alberta
Canada
T9C 1T4

This publication may be cited as:

Florence, L.Z. and L.A. Goonewardene (editors). 1992. Proceedings of the Second All-Alberta Applied Statistics and Biometrics Workshop, October 21-22, 1991. Alberta Environmental Centre, Vegreville, AB. AECV-92P1. 108 pp.

ISBN 0-7732-0912-3

TABLE OF CONTENTS

INTRODUCTION AND ACKNOWLEDGEMENTS

Welcome! This is the second Alberta workshop on applied statistical and biometrical applications. We are again especially impressed by the wide response to our invitation and the diversity of interests. Your active participation will make this year's workshop exceed all expectations, when even compared to last year's success.

We are fortunate to have received the continued support of the Alberta Agricultural Research Institute (AARI) and the AARI Board. Dr. Ralph Christian, Ms. Emily Samoil, and all AARI staff have been willing to assist us in every way.

Special debt is owed many persons who contributed time and effort, often having to work within tight schedules. Mr. Phil Henry, Alberta Environmental Centre (AEC), once again has been responsible for the computer graphics and layouts, registrations and generally keeping the process in control. Mr. John Kirtz, Dr. Alan Tong, Mr. M. Liu and Mr. Serge Dupuis kindly agreed to assist with procuring and demonstrating hardware and software. Dr. Alan Tong, and his staff, are extended special thanks for assisting in the compilation of papers for the proceedings; Ms. Grovenore Flato, AEC, contributed valuable help with wordprocessing and preparation of manuscripts. Our sincere appreciation to Ms. Janet Smalley and the support staff in the Beef Cattle and Sheep Branch of Alberta Agriculture, those in Animal Sciences Division, AEC, and Ms. Gloria Horon, Central Duplicating, AEC, who so kindly met our deadlines on several occasions.

It has been very gratifying to have received the valuable support and participation from those contributing presentations. Through their willingness to share informative topics of interest, and program planning input from Milton Weiss, Alan Tong and Len Kryzanowski, we believe an especially valuable workshop program has again come together. Join us in extending an "Alberta Welcome" to our visiting participants, Drs. George Milliken, Kansas State University and Richard Gilbert, Battele, Northwest Laboratories, Richland, Washington.

We are particularly aware of, and acknowledge, the support we have received from our respective Departments, the Executive Director (AEC), Division Directors and Branch heads. To those we may have mistakenly overlooked, please accept our apologies.

Whether there should be another workshop like this, and last year's, will be a topic for discussion on Tuesday morning (October 22, 0900 h). Regardless what consensus may result, we have been privileged to be associated with many fine people and make new friends, whom we might never have had the opportunity of knowing.

L. Zack Florence
Animal Sciences Division
Biometrics Section
Alberta Environmental Centre

Laki Goonewardene
Animal Industry Division
Beef Cattle and Sheep Branch
Alberta Agriculture

1991 Program

All-Alberta Applied Statistics and Biometrics Workshop

Holiday Inn, Crowne Plaza (formerly Chateau Lacombe) 1011 Bellamy Hill
Edmonton, AB T5J 1N7

Monday, October 21, 1991

All morning sessions are in "Salon B"

Time	Topic
08:00-09:00	Registration (Follow Signs)
09:00-09:15	Welcome and Opening Remarks
09:15-10:00	The Relationship between Methods of Competing Risks Analysis and Methods of Survival Analysis, with Application to Entomology -Bruce Schaalje, Agriculture Canada
10:00-10:15	R & R Break
10:15-11:15	Some Aspects of Analysis of Covariance -George Milliken, Kansas State University
11:15-12:00	Two-Phase Regression - Ray Weingardt, University of Alberta
12:00-12:30	Lunch (Klondike Room)
12:30-14:00	*Graphics and Statistical Software Demonstrations (walk in and try them).

Room	Time	Topic
McDougall	14:00-14:30	Open
	14:30-15:00	Application of the MANOVA to Studies in Agriculture- Laki Goonewardene, Alberta Agriculture
	15:00-15:30	Application of the MANOVA to Studies in Agriculture- Laki Goonewardene, Alberta Agriculture
Strathcona	14:00-14:30	Demonstration of EPI-5 Diagnostic Package in Feedlot Disease Studies-Casey Schipper, Alberta Agriculture
	14:30-15:00	Open
	15:00-15:30	Demonstration of EPI-5 Diagnostic Package in Feedlot Disease Studies-Casey Schipper, Alberta Agriculture
Beaver	14:00-14:30	Open
	14:30-15:00	Neovisuals by SAS (video)
	15:00-15:30	Neovisuals by SAS (video)

1991 Program

All-Alberta Applied Statistics and Biometrics Workshop

**Holiday Inn, Crowne Plaza (formerly Chateau Lacombe) 1011 Bellamy Hill
Edmonton, AB T5J 1N7**

Tuesday, October 22, 1991

All morning sessions are in "Salon B"

Time	Topic
08:00-09:00	Registration (Follow Signs)
09:00-09:45	Workshop 1992? - L.Z. Florence and Laki Goonewardene
09:45-10:15	R & R Break
10:15-11:00	Environmental Restoration and Statistics: Issues and Needs- Richard Gilbert, Battelle, Northwest Laboratories
11:00-11:45	Some Problems in Statistical Consulting -Gerry Kozub, Agriculture Canada
11:45-12:30	Lunch and drawing for Door Prizes (must be present to win!)- Klondike Room

Room	Time	Topic
Klondike	12:30-14:00	*Graphics and Statistical Software Demonstrations
McDougall	14:00-14:30	Running SAS under OS/2 -Serge Dupuis, PWSS, Software Support
	14:30-15:00	Open
	15:00-15:30	Running SAS under OS/2 -Serge Dupuis, PWSS, Software Support
Strathcona	14:00-14:30	Open
	14:30-15:00	Analysis of Air Monitoring Data- Zack Florence and Henry Bertram, Alberta Environmental Centre
	15:00-15:30	Analysis of Air Monitoring Data- Zack Florence and Henry Bertram, Alberta Environmental Centre

Proceedings

The Relationship Between Methods of Competing Risks Analysis and Methods of Survival Analysis, With an Application to Entomology

G. B. Schaalje
Research Station, Agriculture Canada
Lethbridge, Alberta T1J 4B1

Introduction

Time-to-failure data arise in many fields of application including medicine (survival times and remission times of patients receiving various treatments), industry (lifetime testing of products and components), plant science (days to germination or heading of various varieties), and entomology (development times under different temperature regimes). 'Time to failure' is used here as a generic term because a 'failure' could be the achievement of any kind of endpoint, and 'time' could be any kind of non-negative quantity (eg. distance). A unique aspect of such data is that some of the observations may be 'censored'. That is, for some reason not related to the treatment, an individual may not reach the failure endpoint before either the study is terminated or the individual is lost to followup. The time-to-censoring data still contain information about the failure process, and a characteristic of appropriate analysis methods is that they account for this information in addition to the information contained in the noncensored data.

This paper primarily discusses the analysis of time-to-failure data when there are two or more competing causes of failure, usually referred to as competing risks analysis. Methods for analysing time-to-failure data when there is only one cause of failure (survival analysis) are reviewed, and it is demonstrated that these methods are surprisingly difficult to extend to the case of multiple causes of failure. The cautions that must be observed when analysing data with two or more causes of failure are discussed, and an example from entomology is presented. Software for both survival analysis and competing risks analysis is briefly discussed as well.

Survival Analysis for One Failure Type

The goals of survival analysis are to characterize the failure process in useful ways, and to infer the effects of treatments and other covariates on failure. The various methods of survival analysis are centered around the survivor function [S(t)], the density function [f(t)], or the hazard function [h(t)] associated with the failure times.

Survivor Function Methods

The S(t) gives the probability that, for a specified time, an individual's time to failure (survival time) is at least that long. Methods for estimating S(t) for a homogeneous group of individuals are intuitive and simple, the nonparametric product-limit (Kaplan-Meier) and actuarial estimators being the most common. Statistics such as the logrank statistic are available for testing the equality of S(t) for two or more groups. As long as there are only a few discrete-valued covariates, stratification of the data can be used to adjust for covariates. The SAS program LIFETEST and the BMDP program 1L are useful in estimating S(t) and computing related statistics.

Density Function Methods

The use of f(t) as the basis for data analysis is common when there is reason to believe that a parametric distribution such as the Weibull or Lognormal is a good model for the failure times. The advantage is that the data can be summarized in a few fitted parameter values. Maximum likelihood methods are used to fit these models to failure time data, and two or more treatment groups can be compared using likelihood ratio statistics. Another advantage is that regression models for failure time data with discrete or continuous covariates, usually called 'accelerated failure time models', can be formulated in terms of parametric density functions. Although the documentation is a bit confusing, the SAS program LIFEREG and the BMDP program 2L can be used to carry out all of these analyses.

Hazard Function Methods

The h(t) gives the instantaneous rate of failure at a specified time given that the individual has not failed prior to that time, and is related to the risk of failure to which an individual is subject. Nonparametric methods can be used to estimated h(t) for a homogeneous set of individuals. A regression method (proportional hazards or 'Cox' regression) based on h(t) which combines some of the advantages of both nonparametric and parametric analysis has been developed. Under the proportional hazards model, it is assumed that the hazard functions for all individuals are proportional, with the proportionality constant determined by a vector of covariates **x**. The most common model for the covariates is the loglinear model given by

$$h(t;\mathbf{x}) = h_0(t) \exp(\mathbf{b}'\mathbf{x})$$

where $h_0(t)$ is the 'baseline' hazard function for individuals with $\mathbf{x} = \mathbf{0}$. The convenient thing about this model is that $h_0(t)$ contains no information about **b**, and is therefore arbitrary. Thus conditional maximum likelihood can be used to estimate **b** without any restrictions being placed on the form of $h_0(t)$. A unique and useful aspect of this model is that time-varying covariates can be easily incorporated. Also, stratification can be used to accomodate nonproportional hazards within the Cox regression framework in some cases. The SAS program LIFETEST and the BMDP program 1L can be used to obtain a nonsmooth estimate of h(t), and the IMSL subroutine HAZST can be used to obtain a smooth estimate of h(t). Proportional hazards regression can be carried out with the BMDP program 2L and the (soon to be available) SAS program PHREG.

Relationships Among the Functions

All of these analyses are equivalent in the sense any of the three functions can be determined from any other of the three functions upon which the various survival analyses are based. Some of the equations relating the functions are:

$$S(t) = 1 - {}_o\int^t f(s)ds$$

$$S(t) = \exp[- {}_o\int^t h(s)ds]$$

$$h(t) = f(t)/[1 - {}_o\int^t f(s)ds]$$

$$f(t) = h(t) \exp[- {}_o\int^t h(s)ds]$$

$$f(t) = -dS(t)/dt$$

The choice of an analysis method depends on how much is known about the distribution of failure times, on the nature of the covariates, and on the goal of the analysis.

Competing Risks Analysis for Multiple Failure Types

The goals of competing risks analysis are to infer the effects of treatments and covariates on specific causes of failure, to infer relationships among the failure types, and to estimate failure rates due to the remaining causes if some of the causes were to be removed. Methods of competing risks analysis to approach these goals have been developed based on various generalizations of the density function and the hazard function. Generalizations of the survivor function also play a role in some competing risks analyses, but usually as a secondary quantity. Hence, these generalizations will not be discussed here. Two approaches based on different generalizations of the density function, the latent failure time model and the mixture model, will be discussed along with an approach based on generalizations of the hazard function known as cause-specific hazard functions. In this discussion, attention will be restricted to the case of two competing causes of failure.

Latent Failure Time Model

This is the oldest approach to analysis of competing risks, and in some ways would seem to be the most natural generalization of the density function approach for one cause of failure. Under this model it is assumed that each individual possesses latent (or potential) failure times (R and S) for the two causes, with joint density function f(r,s). For convenience it is also assumed that it is impossible to simultaneously fail due to both causes. Since the occurance of a failure due to either of the causes precludes the other, the actual cause of failure is determined by that with the smallest latent failure time. The observable failure time T is then given by

$$T = \min(R,S).$$

If it can be safely assumed that the two latent failure times are independent with density functions $f_R(r)$ and $f_S(s)$, then

$$f(r,s) = f_R(s)\ f_S(s).$$

Various parametric forms have been proposed for f(r,s), and estimation methods, regression models, and methods for evaluating what would happen if some of the causes were to be removed have been developed. There is a major unresolvable problem with this approach, however. Because the latent failure times are not all observable, f(r,s) cannot be estimated without making unverifable assumptions. Any inferences that are made using the latent failure time model also suffer from this difficulty. This whole problem is sometimes referred to as the 'non-identifiability problem' of competing risks analysis.

As a simple example, consider the following hypothetical data:

time	no. living	no. dying (cause R)	no. dying (cause S)
0	1000	0	0
0.5	600	400	-
1	540	-	60
1.5	270	270	-
2	162	-	108
2.5	6696		-
3	13	-	53
3.5	0	13	-

Assuming all events to occur at discrete points in time, a discrete joint density function which fits these data perfectly is:

			S		
R	1	2	3	4	tot
0.5	.040	.144	.173	.043	.400
1.5	.030	.108	.130	.032	.300
2.5	.018	.064	.077	.019	.178
3.5	.012	.044	.053	.013	.122
tot	.100	.360	.433	.107	1.0

However, another joint density function which fits these data perfectly is:

	S				
R	1	2	3	4	tot
0.5	.010	.010	.020	.360	.400
1.5	.010	.010	.010	.250	.280
2.5	.010	.008	.004	.092	.114
3.5	.040	.100	.053	.013	.206
tot	.070	.128	.087	.715	1.0

There are, in fact, an infinite number of joint density functions which fit these data perfectly. If some additional (but unverifiable) assumptions about the form of the distribution of failure times can be made, a unique best-fitting distribution can sometimes be found. For example, if the independence assumption is reasonable, the first joint density function is the unique best-fitting density to these hypothetical data.

My personal feeling is that the latent failure time model should only be used when there is good reason to believe the independence assumption, such as in industrial reliability testing of equipment composed of several independent components. Such independence seems unliklely in most biological applications.

If independence is assumed, the SAS program LIFEREG and the BMDP program 2L can be used to fit the distributions and carry out regression analyses. To use these programs to analyse the latent failure times for one cause, failure times due to the other cause must be treated as censored data.

Mixture Model

In this model, the failure types and times are expressed in terms of a stochastic mechanism which determines the failure type from the outset, and conditional distributions of failure times given the type of failure. The name comes from the fact that the unconditional distribution of failure times is a mixture of the conditional distributions. For two failure types, a binomial distribution [B(n,p)] serves as a model for the mechanism of choosing a failure type, and density functions [$g_R(r)$ and $g_S(s)$] serve as models for the conditional distributions of failure times. This model does not suffer from nonidentifiability problems, and regression models for evaluating the effects of treatments on the failure type mechanism and the conditional failure time distributions have been developed.

The greatest weakness of the mixture model is that although it is identifiable, it may be misleading. Consider, for example, an industrial application in which a piece of equipment with two independent critical components is being tested, so that the independent latent failure time model happens to be valid. Suppose that a treatment is applied which affects only one of the components. It would be desirable if a statistical analysis would indicate that only the one component was being affected by the treatment; the latent failure time model as applied to these data would show just that. The mixture model, however, would very likely indicate that conditional distributions of failure times due to both of the components are affected by the treatment because in fact both conditional distributions would be affected in such a case. The table below gives the true means of the distributions of latent failure times and of the conditional distributions of observed failure times for three specified independent latent failure time processes. Only the distribution of latent failure times for failure type 2 varied among examples, but the means of the conditional distributions for both failure types varied among examples.

Example	Failure Type	Mean of Latent Failure Times	Mean of Conditional Failure Times
1	1	27.2	27.1
	2	36.8	30.5
2	1	27.2	25.6
	2	27.2	25.6
3	1	27.2	24.7
	2	24.7	24.0

If there were no censoring, the model could be fitted in a straightforward way. The binomial p would be estimated by the proportion of individuals subject to failure type 1, and the observed failure times for each failure type would be used in ordinary ways to estimate $g_R(r)$ and $g_S(s)$. If there is censoring, the EM algorithm must be used for estimation and a custom computer program is required.

Câuse-Specific Hazard Model

This model is based on instantaneous failure rate functions [$h_1(t)$ and $h_2(t)$] for each of the failure types. These functions can be estimated and analysed in exactly the same way as ordinary hazard functions using the same software. The only modification necessary is that failure times for failures of types other than the one under consideration must be regarded as censored observations. This analysis is not subject to nonidentifiability nor misinterpretation problems, so long as it is kept in mind that conclusions based on these analyses apply only under the prevailing study conditions, with all failure types operative. A further benefit is that relationships among the failure types can be studied via the use of time-varying covariates if 'risk-indicator' variables for some of the failure types are available.

Relationships Among the Models

Relationships among the various competing risks models are discussed by Gail (1975), Gail (1982), and Prentice et al. (1978).

Example

This example involves data on the survival and development times of migratory locusts treated with the pathogens *Nosema locustae* and *Nosema cuneatum.* Although these disease organisms do not inflict as much mortality as would be desired for a control agent, it was believed that they might provide additional crop protection by prolonging development through reduced consumption and activity. The objective of this study was to determine and compare the effects of *N. locustae* and *N. cuneatum* infections, at a number of levels, on development and mortality of locusts.

The study was carried out using individually caged third instar locusts which were innoculated with Nosema spores at one of 6 rates. The two 'failure' types were death and the attainment of the adult stage. The time to reach one of these endpoints was recorded for each locust. The study was stopped after 36 days, and any locusts still alive and in a subadult stage were considered to be censored. All three competing risks models were applied to the data.

1. The mixture model

This model was applied to the locust data simply by computing the proportions of individuals dying or reaching adulthood, and examing the observed (conditional) distributions of death and development times (see table below). This is appropriate for the N. locustae data because there was virtually no censoring, but is not a full fitting of the model for the N. cuneatum data. This approach corresponds to the customary analysis of such data.

The mean observed development and death times were fairly constant over all dosage levels of spores of both *Nosema* species. If it were not known that conditional development times can be misleading about the effects of a treatment on competing failure types, one would likley conclude that *Nosema* does not affect development. The proportion of locusts completing development decreased with increasing spore dose, but these proportions do not tell much about the impact of *Nosema* on development because such a decrease could simply have been due to the increased mortality with increasing spore doses.

			Percent		Ave. Time to	
Species	Spore Dosage	Num.	Mort.	Devel.	Mort.	Devel.
locustae	0	81	4	96	26.3	27.5
	10^2	82	2	98	23.5	26.6
	10^3	77	39	60	29.0	28.4
	10^4	75	72	28	26.8	26.6
	10^5	89	90	9	27.8	29.0
	10^6	88	99	1	26.2	27.0
cuneatum	0	75	1	99	20.0	28.1
	10^2	90	3	96	19.0	28.7
	10^3	75	3	96	26.5	27.5
	10^4	72	32	57	28.1	29.3
	10^5	87	59	11	27.7	29.7
	10^6	86	77	5	25.7	31.0

2. The latent failure time model

As can be seen in the following table, the estimated means of the distributions of latent development and death times showed trends with increasing spore dose.

Species	Spore Dosage	Ave. Latent Time to Death	Ave. Latent Time to Devel.
locustae	0	40.5	27.6
	10^2	58.1	26.7
	10^3	31.9	30.4
	10^4	27.4	29.5
	10^5	28.2	35.2
	10^6	26.2	36.4
cuneatum	0	117.1	28.1
	10^2	119.2	28.7
	10^3	48.1	27.5
	10^4	34.5	31.7
	10^5	32.8	42.0
	10^6	28.8	41.9

If the assumption of independence of the latent times were true, these results would indicate that as spore dose increases, expected survival time of the locusts decreases, especially for *N. locustae*. They would also indicate that even if all insects were to survive, they would take much longer to complete development to adulthood under the influence of high doses (> 10^4) of *Nosema* spores of either species. There would also be an indication that *N. cuneatum* prolongs development to a greater extent than *N. locustae*.

3. The cause-specific hazard model

The estimated development hazard and death hazard functions demonstrated conclusively that both the mortality and development processes were affected by *Nosema*. The death hazard function was nonzero for *N. locustae* doses greater than or equal to 10^3 and for *N. cuneatum* doses greater than or equal 10^4. Analysis of the death hazard functions via the proportional hazards model gave the following proportionality factors for doses greater than or equal to 10^4.

Spore Dosage	Proportionality Factor	
	locustae	*cuneatum*
10^4	1.00	0.25
10^5	0.95	0.39
10^6	1.51	0.72

The development hazard function seemed unaffected by *N. locustae* at a dose of 10^2 spores, but for doses of 10^3 spores or more the hazard function decreased toward 0 in a regular manner. For *N. cuneatum*, the development hazard function was unaffected for doses less than or equal to 10^3, but was just as depressed as for *N. locustae* at higher doses. Proportional hazards analysis of the development hazard functions (see table below) showed that the effects of equivalent doses of *N. locustae* and *N. cuneatum* on development were very similar except at the lowest doses, 10^2 and 10^3 spores.

Spore Dosage	Proportionality Factor	
	locustae	*cuneatum*
0	1.00	0.89
10^2	1.40	0.74
10^3	0.42	1.02
10^4	0.42	0.32
10^5	0.08	0.06
10^7	0.02	0.03

The death hazard functions were reanalysed using a proportional hazards model with a time-dependent covariate. The value of this covariate was 0 (zero) if the locust had not yet achieved the fifth instar, and was 1 after it became a fifth-instar locust. Obviously, this covariate contains crude information about the nearness of an individual to the adult stage, and can be used to investigate the relationship between the two competing endpoints. The following proportionality factors were obtained:

Spore Dosage	Age	Proportionality Factor	
		locustae	*cuneatum*
10^4	inst<5	1.00	0.88
	inst=5	0.79	0.15
10^5	inst<5	0.87	0.95
	inst=5	0.69	0.16
10^6	inst<5	1.51	1.30
	inst=5	1.19	0.22

This analysis shows that for locusts treated with *N. cuneatum*, there was a large reduction in the death hazard upon the onset of the fifth instar, whereas for those treated with *N. locustae* the reduction was very small. Thus for *N. cuneatum*, the competing endpoints strongly influence each other.

This analysis also suggests that the modes of action of the two species of *Nosema* are quite different. A related interesting result of this analysis is that when the time-dependent covariate is in the model, there is no longer a significant difference between the species of *Nosema*. Apparently most of the difference between species in their death hazard functions can be explained by the reduction in the death rate which occurs at the onset of the fifth instar for locusts treated with *N. cuneatum*. The use of the cause specific hazard model has provided information that no other procedure would have detected, and which could not have been even guessed at by a superficial summary of the data.

References

BMDP INC. 1990. BMDP statistical software manual. University of California Press, Berkeley.

COX, D.R. 1972. Regression models and life tables (with discussion). J. Royal Stat. Soc., Series B 34: 187-220.

COX, D.R. AND D. OAKES. 1984. Analysis of Survival Data. Chapman and Hall, London.

DAVID, H.A., AND M.L. MOESCHBERGER. 1978. The theory of competing risks. Griffin, High Wycombe.

ELANDT-JOHNSON, R.C., AND N.L. JOHNSON. 1980. Survival Models and Data Analysis. Wiley, New York.

GAIL, M. 1975. A review and critque of some models used in competing risks analysis. Biometrics 31: 209-222.

GAIL, M. 1982. Competing risks, pp. 75-81 In S. Kotz and N.L. Johnson (eds.), Encyclopedia of statistical sciences, vol. 2. Wiley, New York.

HECKMAN, J.J., AND B.E. HONORE. 1989. The identifiability of the competing risks model. Biometrika 76: 325-330.

IMSL INC. 1989. User's manual IMSL STAT/LIBRARY. IMSL Inc., Houston.

KALBFLEISCH, J.D., AND R.L. PRENTICE. 1980. The statistical analysis of failure time data. Wiley, New York.

LARSON, M.G., AND G.E. DINSE. 1985. A mixture model for the regression analysis of competing risks data. Appl. Statist. 34: 201-211.

LAWLESS, J.E. 1982. Statistical Models and Methods for Lifetime Data. Wiley, New York.

MILLER, R.G. Jr. 1981. Survival Analysis. Wiley, New York.

PRENTICE, R.L., J.D. KALBFLEISCH, A.V. PETERSON Jr., N. FLOURNAY, V.T. FAREWELL, AND N.E. BRESLOW. 1978. The analysis of failure times in the presence of competing risks. Biometrics 34: 541-554.

SAS INSTITUTE INC. 1989. SAS/STAT user's guide, version 6, fourth edition, vol. 2. SAS Institute Inc., Cary NC.

TSIATIS, A. 1975. A nonidentifiability aspect of the problem of competing risks. Proc. Nat. Acad. Sci. 72: 20-22.

YASHIN, A.I., K.G. MANTON, AND E. STALLARD. 1986. Dependent competing risks: a stochastic process model. J. Math. Biol. 24: 119-164.

Some Aspects of Analysis of Covariance

George A. Milliken
Department of Statistics, Dickens Hall
Kansas State University
Manhattan, KS 66506

ABSTRACT

The statistical procedure, analysis of covariance has been used in several contexts. The most common description of analysis of covariance is to adjust the analysis for variables that could not be controlled by the experimenter. For example, a researcher can remove the differential effects of a fertility trend by using a randomized complete block design structure, but it may not be possible to control the number of plants per plot of land. The researcher wishes to compare varieties as if each plot had the same number of plants. The analysis of covariance is a procedure which can compare variety means by first adjusting for the differential number of plants per plot.

The analysis of covariance described here is in a more general context than that of adjusting for variation due to uncontrollable variables. The ANALYSIS OF COVARIANCE is defined as a method for comparing several regression surfaces or lines, one for each treatment or treatment combination, where there is possible a different regression surface for each treatment or treatment combination.

1. Introduction

The experimental situation involves randomly assigning n_i experimental units to treatment i, applying the treatments and measuring $y_i, x_{1ij}, x_{2ij}, \ldots, x_{qij}$ on each experimental unit, where

y_{ij} is the dependent measure,

x_{1ij} is the first independent variable or covariate,

x_{2ij} is the second independent variable or covariate,

x_{kij} is the kth independent variable or covariate.

At this point, the experimental design is a one-way treatment structure in a completely randomized design structure with k covariates. The covariance model consisting of a linear function of the covariates or independent variables is

$$y_{ij} = \beta_{oi} + \beta_{1i}x_{1ij} + \beta_{2i}x_{2ij} + \ldots + \beta_{ki}x_{kij} + \varepsilon_{ij} \qquad (1.1)$$

for i=1,2,...t where t is the number of treatments, $j = 1,2,\ldots,n_i$, and the $\varepsilon_{ij} \sim N(0, \sigma^2)$. The important thing to note about this model is the mean of the y values for given values of the x's depends on the values of the x's as well as on the treatment or treatment combination or population from which the data were collected. The analysis of covariance is a strategy for making decisions about the form of the covariance model through testing hypotheses and then comparing the treatments by comparing the estimated responses from the resulting regression models.

Possible forms of hypotheses are:

H_{o1}: $\beta_{h1} = \beta_{h2} = ... = \beta_{ht} = 0$ vs H_{a1}: (not H_{o1}:), that is, all the slopes for the hth covariate are zero, or

H_{o2}: $\beta_{h1} = \beta_{h2} = ... = \beta_{ht}$ vs H_{a2}: (not H_{o2}:), that is, all the slopes for the hth covariate are equal meaning,the surfaces are parallel in the direction of the hth covariate.

The analysis of covariance model is a combination of the analysis of variance model and the regression model. An experiment is designed to purchase a certain number of degrees of freedom for error (generally with out the covariates) and the experimenter is willing to sell some of those degrees of freedom for good covariates which will help reduce the magnitude of the error variance. The philosophy in this book is to select the simplest possible expression for the covariate part of the model before making treatment comparisons. This process of model building to determine the simplest adequate form of the model follows the principle of parsimony as well as helps guard against foolishly selling degrees of freedom for error to obtain unnecessary covariate terms in the model. Thus the strategy for analysis of covariance begins with testing hypotheses like above to make decisions about the form of the covariate or regression part of the model.

The structure of the following chapters leads one through the forest of analysis of covariance by starting with the simple model with one covariate through the complex process involving analysis of covariance in split-plot and repeated measures designs. Other topics discussed are multiple covariates, experiments involving blocks, and graphical methods for comparing the models for the various treatments.

2. One-way Analysis of Covariance--One Covariate in a Completely Randomized Design Structure

2.1 The Model

Suppose there are N homogeneous experimental units which are randomly divided into groups of n_i units per group where

$$\sum_{i=1}^{t} n_i = N.$$

Each of the t treatments of a one-way treatment structure is randomly assigned to a group of experimental units, providing a one-way treatment structure in a completely randomized design structure. Let y_{ij} (dependent variable) denote the jth observation from the ith treatment and x_{ij} denote the covariate (independent variable) corresponding to the (i,j)th experimental unit. Assume that the mean of y_{ij} can be expressed as a linear function of the covariate with possibly different linear functions required for each treatment. It is important to note that the mean of an observation from the ith treatment group, depends on the particular treatment as well as on the values of the covariate (independent variable).

The analysis of covariance model for a one-way treatment structure with one covariate in a completely randomized design structure is

$$y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij},$$
$$i = 1,2,\ldots,t. \;\; j = 1,2,\ldots,n \tag{2.1}$$

where the mean of y_i for a given value of X is $\mu_{Y_i|X} = \alpha_i + \beta_i X$ and for making inferences, it is assumed that

$$\varepsilon_{ij} \sim \text{iid } N(0,\sigma^2).$$

Model (2.1) has t intercepts, $\alpha_1,\ldots,\alpha_t$, t slopes $\beta_1,\ldots,\beta_t$ and one variance σ^2. Before one can continue with the analysis of this model, one must be sure that the data from each treatment can in fact be described by a simple linear regression model. Various regression diagnostics should be run on the data before continuing. The equal variance assumption should also be checked. If the simple linear regression model is not adequate to describe the data for each treatment another model must be selected before continuing with the analysis of covariance. The analysis of covariance involves comparing the t slopes and comparing the distances between the regression lines (surfaces) at preselected values of X. The analysis of covariance computations are typically presented in summation notation with little emphasis on interpretations. In this paper the various covariance models are expressed in terms of matrices (see Chapter 6 of Milliken and Johnson (1984)) and their interpretations are discussed. The computer is used as the mode of doing the analysis of covariance computations.

The matrix form of Model (2.1) is

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{t1} \\ \vdots \\ y_{tn_t} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & x_{21} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{t1} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & x_{tn_t} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \\ \beta_1 \\ \alpha_2 \\ \\ \beta_2 \\ \vdots \\ \alpha_t \\ \\ \beta_t \end{bmatrix} + \underline{\varepsilon}. \tag{2.2}$$

which is expressed in the form of a linear model as $\underline{y} = \underline{X}\beta + \underline{\varepsilon}$. The vector $\underline{\beta}$ denotes the collection of slopes and intercepts, the matrix $\underline{X}$ is the design matrix and the vector $\underline{\varepsilon}$ represents the random errors.

2.2 Estimation

The least squares estimator of the parameter vector $\underline{\beta}$ is

$$\underline{\beta} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}.$$

But the least squares estimator of β can also be found by obtaining the least squares estimator of each pair of parameters (α_i, β_i) by fitting the simple linear model to the data from each treatment. For data from the ith treatment, fit the model

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix} \begin{bmatrix} \alpha_i \\ \\ \beta_i \end{bmatrix} + \underline{\varepsilon}_i, \tag{2.3}$$

which is expressed as $\underline{y}_i = \underline{X}_i\underline{\beta}_i + \underline{\varepsilon}_i$. The least squares estimator of $\underline{\beta}_i$ is

$$\underline{\beta}_i = (\underline{X}_i'\underline{X}_i)^{-1}\underline{X}_i'\underline{y}_i,$$

the same as the estimator obtained for a simple linear regression model. The estimates of β_i and α_i in summation notation are

$$\beta_i = \frac{\sum_{j=1}^{n_i} x_{ij}y_{ij} - n_i\bar{x}_{i.}\bar{y}_{i.}}{\sum_{j=1}^{n_i} x_{ij}^2 - n_i\bar{x}_{i.}^{2}}$$

and

$$\hat{\alpha}_i = \bar{y}_{i.} - \beta_i\bar{x}_{i.}.$$

The residual sum of squares for the ith model is

$$SSRes_i = \sum_{j=1}^{n_i} (y_{ij} - \hat{\alpha}_i - \beta_i x_{ij})^2$$

which is based on n_i-2 degrees of freedom. After testing the equality of the treatment variances, the residual sum of squares for model (2.1) can be obtained by pooling residual sums of the squares for each of the t models, i.e., sum them together to obtain

$$SSRes = \sum_{i=1}^{t} SSRes_i. \tag{2.4}$$

The pooled residual sum of squares, SSRes, is based on the pooled degrees of freedom

$$\text{d.f.}_{\text{SSRes}} = \sum_{i=1}^{t} (n_i - 2) = \sum_{i=1}^{t} n_i - 2t = N - 2t.$$

The best estimate of the experimental unit variance is

$$\hat{\sigma}^2 = \text{SSRes}/(N - 2t).$$

The sampling distribution of

$$(N - 2t)\hat{\sigma}^2/\sigma^2$$

is central chi-square with (N - 2t) degrees of freedom. The sampling distribution

estimator,

$$\hat{\underline{\beta}}' = (\hat{\alpha}_1, \hat{\beta}_1, \ldots, \hat{\alpha}_t, \hat{\beta}_t)$$

is normal with mean $\underline{\beta}' = (\alpha_1,\beta_1,\ldots,\alpha_t,\beta_t)$ and variance-covariance matrix $\sigma^2(\underline{X}'\underline{X})^{-1}$, which can be written as

$$\sigma^2(\underline{X}'\underline{X})^{-1} = \sigma^2 \begin{bmatrix} (\underline{X}_1'\underline{X}_1)^{-1} & & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (\underline{X}_t'\underline{X}_t)^{-1} \end{bmatrix} \tag{2.5}$$

where

$$\underline{X}_i = \begin{bmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix}$$

2.3 Strategy for Determining the Form of the Model

The main objective of an analysis of covariance is to compare the t regression lines at several predetermined fixed values of the covariate, X. Depending on the values of the β_i, there are various strategies that one takes when comparing the regression lines.

The first question which needs to be answered is, does the mean of y given X depend on the value of X? That question can be answered "statistically" by testing the hypothesis H_{01}: $E(y_{ij} \mid X = x) = \alpha_i$ vs. H_{a1}: $E(y_{ij} \mid X_{ij} = x_{ij}) = \alpha_i + \beta_i x$ for i = 1,2,...,t. The hypothesis is equivalent to

$$H_{01}:\ \beta_1 = \beta_2 = \ldots = \beta_t = 0 \ \text{ vs. } \ H_{a1}:\ (\text{not } H_0). \tag{2.6}$$

The null hypothesis states that none of the treatments have means which depend linearly on the value of the covariate, X.

The principle of conditional error (Milliken and Johnson, (1984)) or model comparison method (Draper and Smith (1981)) provides an excellent way of obtaining the desired test statistic. The model restricted by the conditions of the null hypothesis, H_{01}, is

$$y_{ij} = \alpha_i + \varepsilon_{ij} \; i = 1,2,\ldots,t, \; j = 1,2,\ldots,n_i. \tag{2.7}$$

Model (2.7) is the usual analysis of variance model for the one-way treatment structure in a completely randomized design structure. The residual sum of squares for model (2.7) is

$$\text{SSRes}(H_{01}) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \tag{2.8}$$

which is based on

$$\text{d.f.}_{\text{SSRes}(H_{01})} = N - t \text{ degrees of freedom}$$

(where the mean of the model under H_{01} has t parameters). The sum of squares due to deviations from H_{01}, denoted by SSH_{01} is,

$$SSH_{01} = \text{SSRes}(H_{01}) - \text{SSRes}, \tag{2.9}$$

which is based on

$$\text{d.f.}_{\text{SSRes}(H_{01})} - \text{d.f.}_{\text{SSRes}} = (N - t) - (N - 2t) = t \text{ degrees of freedom.}$$

The sampling distribution of SSH_{01}/σ^2 is a noncentral chi-square distribution with t degrees of freedom where the non-centrality parameter is zero if and only if H_{01} is true. A statistic for testing H_{01} versus H_{a1} is

$$F_{H_{01}} = \frac{SSH_{01}/t}{\hat{\sigma}^2} \tag{2.10}$$

and, when H_{01} is true, the sampling distribution of $F_{H_{01}}$ is a central F distribution with t and N - 2t degrees of freedom.

If one fails to reject H_{01}, then one can conclude that the means of the treatments do not depend linearly on the value of the covariate, X. In this case, the next step in the analysis is to use analysis of variance to make comparisons between treatment means, i.e., compare the α_i, i = 1,2,...,t (see chapter 1, Milliken and Johnson (1984)). Recall it has already been determined that the simple linear regression model adequately describes the data. Thus if the slopes are zero, then conclude the models are of the form $y_{ij} = \alpha_i + \varepsilon_{ij}$.

If H_{01} is rejected, then conclude that the mean of y does depend linearly on the value of the covariate X for at least one of the treatments. In this case, the next step in the analysis of covariance is to determine whether or not the means of the treatments depend on the covariate X differently (as represented by unequal slopes which provide non-parallel lines). A test for homogeneity or equality of the slopes will answer that question. The appropriate null hypothesis is

$$H_{02}\text{: } E(y_{ij} \mid X = x) = \alpha_i + \beta x \quad \text{vs.} \quad H_{a2}\text{: } E(y_{ij} \mid X = x) = \alpha_i + \beta_i x \text{ or equivalently}$$

$$H_{02}\text{: } \beta_1 = \beta_2 = \ldots = \beta_t = \beta \quad \text{vs.} \quad H_{a2} \text{ (not } H_{02}) \tag{2.11}$$

where β is unspecified and represents the common slope of the t parallel regression lines. The model in the form of (2.1) which satisfies the conditions of H_{02} is

$$y_{ij} = \alpha_i + \beta x_{ij} + \varepsilon_{ij} \quad i = 1,2,...,t, \;\; j = 1,2,...,n_i \tag{2.12}$$

which represents t parallel lines each with slope β and intercepts $\alpha_1,...,\alpha_t$. The matrix form of model (2.12) is

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{t1} \\ \vdots \\ y_{tn_t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & x_{11} \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & 0 & X_{1n_1} \\ 0 & 1 & \cdots & 0 & x_{21} \\ \vdots & \vdots & & & \vdots \\ 0 & 1 & \cdots & 0 & x_{2n_2} \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 & x_{t1} \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 1 & x_{tn_t} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_t \\ \vdots \\ \beta \end{bmatrix} + \underline{\varepsilon} \,. \tag{2.13}$$

The residual sum of squares for model (2.13) is

$$\text{SSRes }(H_{02}) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \hat{\alpha}_i - \hat{\beta} x_{ij})^2 \tag{2.14}$$

where

$$\hat{\alpha}_i \;\; i = 1,2,...,t \text{ and } \hat{\beta}$$

denote the least squares estimators of the corresponding parameters from model (2.13). The residual sum of squares in (2.14) is based on

$$\text{d.f.}_{\;\text{SSRes}(H_{02})} = \text{N-t-1 degrees of freedom}$$

as the mean of model (2.12) has t + 1 parameters. The sum of squares due to deviations from H_{02} is

$$SSH_{02} = \text{SSRes}(H_{02}) - \text{SSRes} \tag{2.15}$$

which is based on

$$\text{d.f.}_{\;\text{SSRes}(H_{02})} - \text{d.f.}_{\;\text{SSRes}(H_{02})} = \text{t - 1 degrees of freedom.}$$

The sampling distribution of SSH_{02}/σ^2 is non-central chi-square with t - 1 degrees of freedom where the non-centrality parameter is zero if and only if H_{02} is true. The statistic used to test H_{02} is

$$F_{H_{02}} = \frac{SSH_{02}/t - 1}{\hat{\sigma}^2} \tag{2.16}$$

which has sampling distribution F with t - 1 and N - 2t degrees of freedom. If one fails to reject H_{02}, then conclude that the lines are parallel (equal slopes) and proceed to compare the distances between the parallel regression lines by comparing their intercepts, α_i's (which is discussed in Section 2.4). Figure 2.1 displays the relationships between the treatment means as a function of the covariate X when the lines are parallel. Since the lines are parallel, i.e., the distance between any two lines is the same for all values of X, a comparison of the intercepts is a comparison of the distances between the lines.

If one rejects H_{02}, then conclude that at least two of the regression lines have unequal slopes and hence, the set of lines are not parallel. Figure 2.2 displays a possible relationship between the means of treatments as a linear function of the covariate for the non-parallel line case. When the lines are not parallel, the distance between two lines depends on the value of X, thus nonparallel line case is called covariate by treatment interaction.

2.4 Comparing the Treatments or the Regression Lines

The correct method for comparing the distances between the regression lines depends on the decision one made concerning the slopes of the models. If the experimenter rejects H_{01} in (2.6) and fails to reject H_{02} in (2.11), the resulting model is a set of parallel lines (equal slopes). A property of two parallel lines is that they are the same distance apart for every value of X. Thus, the distance between any two lines can be measured by comparing the intercepts of the two lines. When the lines are parallel, contrasts between the intercepts are used to compare the treatments. When the slopes are unequal there are two types of comparisons that are of interest to the experimenters, namely, comparing the distances between the various regression lines at several values of the covariate X and comparing specific parameters, such as slopes or the models evaluated at the same or different values of X.

2.4.1 Equal Slope Model

At this step in the analysis remember that H_{02} was not rejected, thus the model used to describe the mean of y as a function of the covariate is of the form

$$y_{ij} = \alpha_i + \beta X_{ij} = \varepsilon_{ij}. \tag{2.17}$$

The residual sum of squares for model (2.17) is $SSRes(H_{02})$ which was given in equation (2.13). The first hypothesis to be tested is that the distances between the lines are equal, which is equivalent to testing the hypothesis that the intercepts are equal as

$$H_{03}\text{:}\ \ \alpha_1 = \alpha_2 = \ldots = \alpha_t = \alpha \ \text{ vs. }\ H_{a3}\text{:}\ \ (\text{not } H_0) \tag{2.18}$$

where α is unspecified. The model in (2.17) restricted by the conditions of H_{03} is

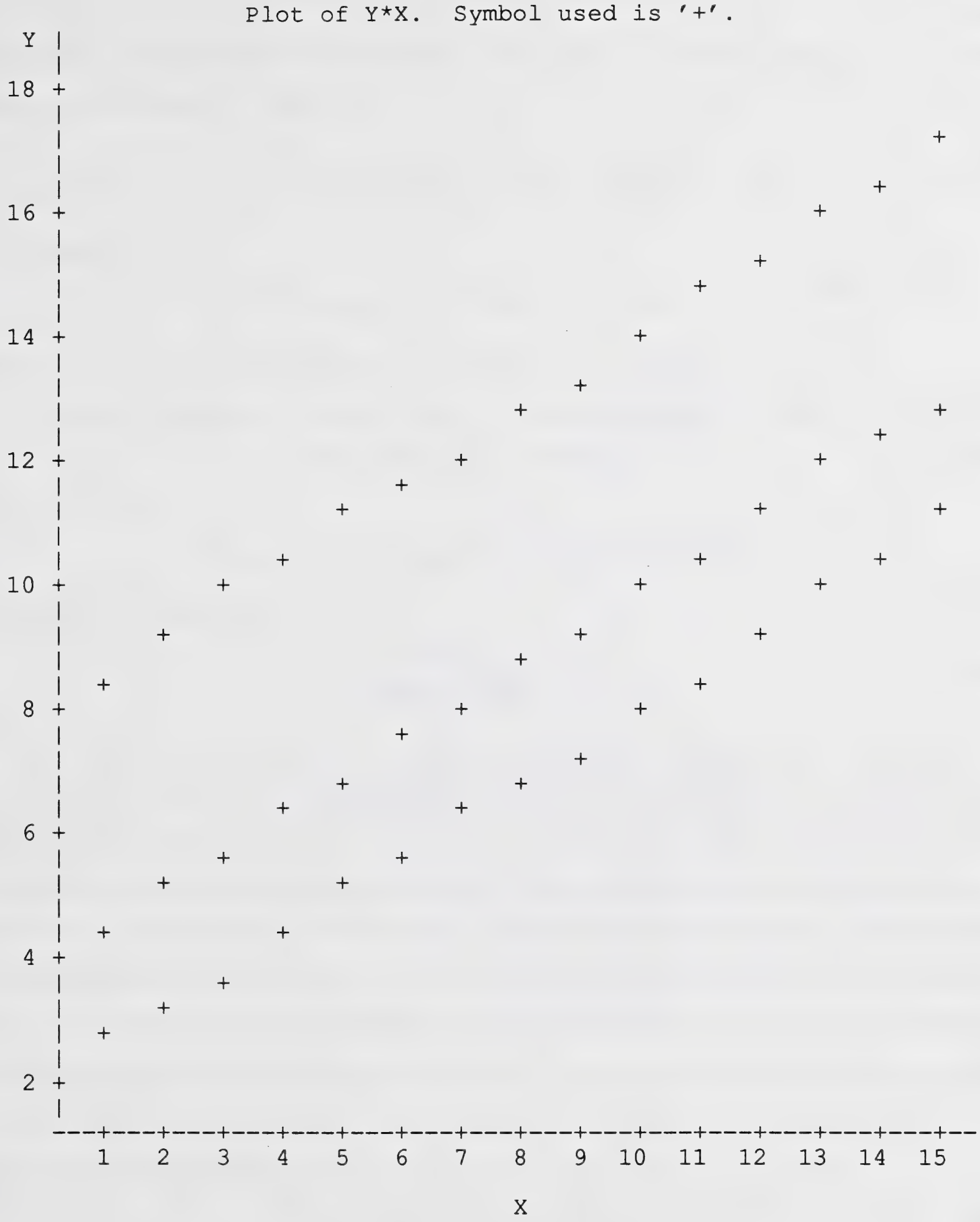


Figure 2.1   Graph of parallel lines models---common slopes.

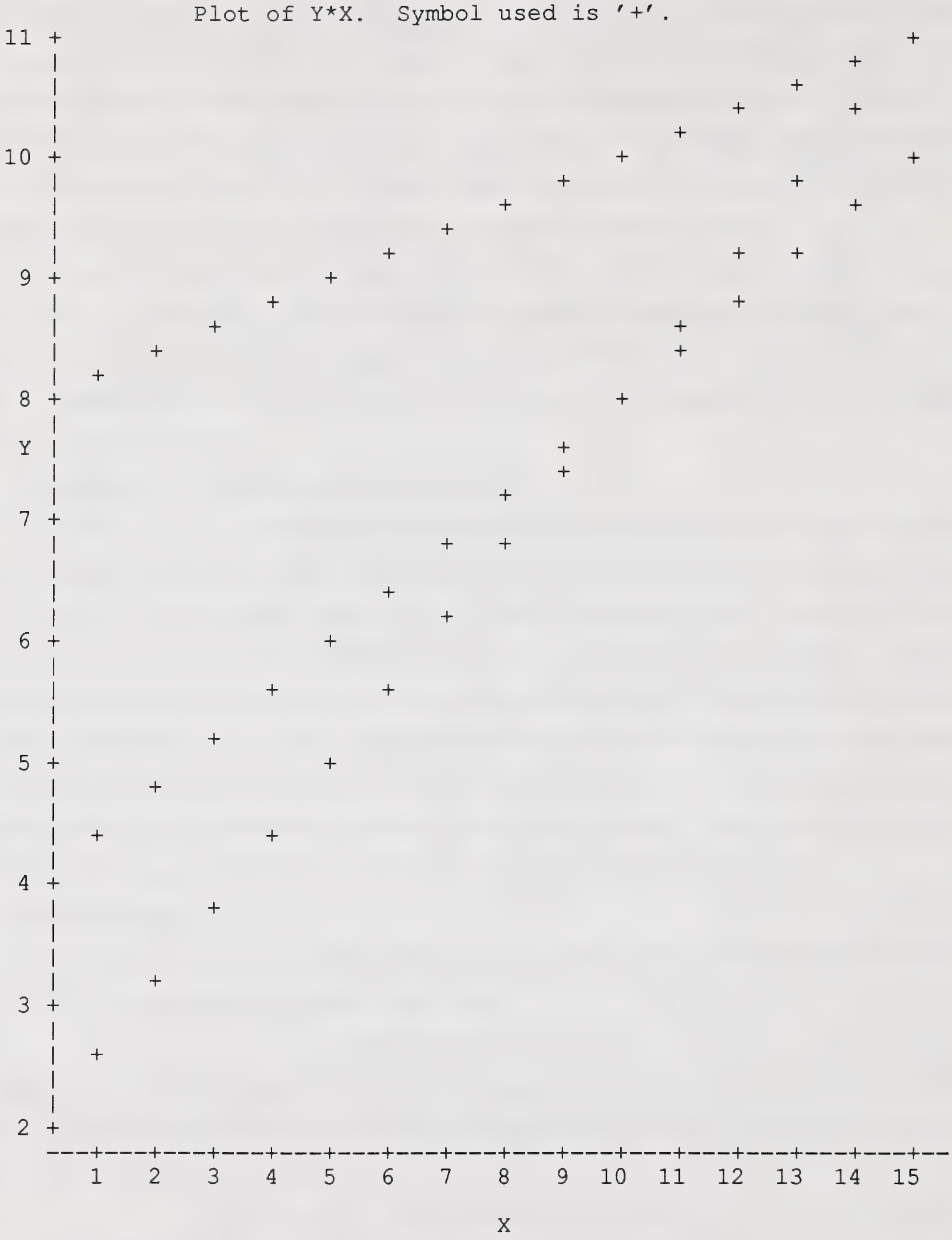


Figure 2.2 Graph of non-parallel lines models---unequal slopes.

$$y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij}. \tag{2.19}$$

Model (2.19) is a single simple linear regression model which is to be fit to all of the data and the corresponding residual sum of squares is

$$\text{SSRes}(H_{03}) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \hat{\alpha} - \hat{\beta} X_{ij})^2 \tag{2.20}$$

which is based on

$$\text{d.f.}_{\text{SSRes}(H_{03})} = N - 2 \text{ degrees of freedom}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are least squares estimators of α and β from model (2.19).

The sum of squares due to deviation from H_{03}, given that the slopes are equal (H_{02} is true), is

$$\text{SSH}_{03} = \text{SSRes}(H_{03}) - \text{SSRes}(H_{02}) \tag{2.21}$$

which is based on

$$\text{d.f.}_{\text{SSRes}(H_{03})} - \text{d.f.}_{\text{SSRes}(H_{02})}) = t - 1 \text{ degrees of freedom.}$$

The appropriate test statistic is

$$F_{H_{03}} = \frac{\text{SSH}_{03}/(t - 1)}{\text{SSRes}(H_{02})/(N - t - 1)} \quad . \tag{2.22}$$

The sampling distribution of $F_{H_{03}}$ is that of a non-central F distribution with t - 1 and N - t - 1 degrees of freedom. If H_{03} is not rejected conclude that all of the data comes from a single regression model with slope β and intercept α, i.e., there are no treatment differences. If H_{03} is rejected, then conclude that the distances between one or more pairs of lines are different from zero. Since the lines are parallel, the distance between any two lines can be compared at any chosen value of X. If the distance between any two lines is compared at X = 0, it is a comparison between the difference of two intercepts as $\alpha_i - \alpha_{i'}$. A multiple comparison procedure can be used to compare the distances with an LSD type being used for controlling the comparison wise error rate and a Bonferroni or Scheffe type being used to control the experiment wise error rate. To use a multiple comparison procedure, one must compute the standard error of the difference between two α's as

$$S_{\hat{\alpha}_i - \alpha_{i'}} = \left[S^2_{\hat{\alpha}_i} + S^2_{\hat{\alpha}_{i'}} - 2\text{Cov}(\hat{\alpha}_i, \hat{\alpha}_i'\right]^{1/2} \tag{2.23}$$

where

$S^2_{\hat{\alpha}}$ is the variance of $\hat{\alpha}_i$ and $\text{Cov}(\hat{\alpha}_i, \hat{\alpha}_{i'})$ is the covariance between $\hat{\alpha}_i$ and $\hat{\alpha}_{i'}$.

If there are specific planned comparisons (such as linear or quadratic effects for treatments with quantitative levels) between the treatments, those comparisons would be made by constructing the necessary contrasts between the intercepts.

When analysis of covariance was first developed, it was mainly used to adjust the mean of y for a selected value of the covariate. The value usually selected was the mean of the covariate from all t treatments. Thus the term adjusted means was defined as the mean of y evaluated at

$$X = \bar{x}\text{...}, \text{ where } \bar{x}\text{..}$$

is the mean value of all the x_{ij}'s. The estimators of the means of treatments at

$$X = \bar{x}\text{..},$$

called the adjusted means, are

$$\hat{\mu}_{Y_i|\beta X = \beta\bar{x}} = \hat{\alpha}_i + \hat{\beta}\bar{x}\text{.. } i = 1,2,\ldots,t., \tag{2.24}$$

where $\hat{\alpha}_i$ and $\hat{\beta}$ are least squares estimators of α_i and β from model (2.18). The covariance matrix of the adjusted means can be constructed from the elements of the covariance matrix of $\hat{\alpha}_1,\ldots.\hat{\alpha}_t$ and $\hat{\beta}$.

The standard errors of the adjusted means are computed as

$$S_{\hat{\mu}_{Y_i|\beta X = \beta\bar{x}}} = \left[S^2_{\hat{\alpha}_i} + \bar{X}^2 S^2_{\hat{\beta}} + 2\bar{X}\text{cov}(\hat{\alpha}_i,\hat{\beta})\right]^{1/2}, \; i = 1,2,\ldots,t.$$

One hypothesis of interest is that of the equality of the adjusted means. This hypothesis can be expressed as

$$H_{04}: \; \mu_{Y_1|\beta X = \beta\bar{x}} = \ldots = \mu_{Y_t|\beta X = \beta\bar{x}} \text{ vs. } H_a(\text{not } H_{04}).$$

But since the lines are parallel, the difference between two adjusted means is the difference between intercepts as

$$\mu_{Y_1|\beta x = \beta\bar{x}} - \mu_{Y_2|\beta X = \beta\bar{x}} = \alpha_1 - \alpha_2,$$

thus H_{04} is equivalent to H_{03}.

Preplanned treatment comparisons and multiple comparison procedure can be carried out to compare the adjusted means by computing the standard error of the difference between pairs of adjusted means. Since the difference of two such means is $\hat{\alpha}_i - \hat{\alpha}_{i'}$, the standard error in (2.23) can be used with any selected multiple comparison procedure. Contrasts between adjusted means, which are also contrasts between the intercepts, measuring linear, quadratic, etc trends should be used when appropriate.

2.4.2 Unequal Slope Model-Covariate by Treatment Interaction

When H_{02} is rejected, it is concluded that the non-parallel lines model (2.1) is necessary to adequately describe the data. The graph of such a possible situation was given in figure (2.2). Since the lines are not parallel, the distance between any two lines depends on which value of the covariate is selected. This is called covariate by treatment interaction. In the nonparallel lines case a comparison of the intercepts is only a comparison of the lines at X = 0. That will generally be a meaningful comparison only when X = 0 is included in or is close to the range of X values in the experiment. The equal intercept hypothesis given the slopes are unequal is expressed as

$$H_{05}\text{: } E(y_{ij} \mid X = x_{ij}) = \alpha + \beta_i x_{ij} \text{ vs } H_{a5}\text{: } E(y_{ij} \mid X = x_{ij}) = \alpha_i + \beta_i x_{ij} \text{ or equivalently,}$$

$$H_{05}\text{: } \alpha_1 = \alpha_2 = ... = \alpha_t = \alpha \text{ given } \beta_i \neq \beta_{i'} \text{ vs } H_a(\text{not } H_0).$$

The model comparison method is used to test this hypothesis. Model (2.1) restricted by the conditions of H_{05} is

$$y_{ij} = \alpha + \beta_i X_{ij} + \varepsilon_{ij} \tag{2.25}$$

and the corresponding residual sum of squares as

$$SSRes(H_{05}) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \hat{\alpha} - \hat{\beta}_i X_{ij})^2 \tag{2.26}$$

which is based on

$$d.f._{SSRes(H_{05})} = N - t - 1 \text{ degrees of freedom.}$$

The values of $\hat{\alpha}$ and $\hat{\beta}_i$ in (2.26) are the least squares estimators of the parameters of model (2.25). The sum of squares due to deviations from H_{05} is

$$SSH_{05} = SSRes(H_{05}) - SSRes,$$

which is based on

$$d.f._{SSRes(H_{05})} - d.f._{SSRes} = t - 1 \text{ degrees of freedom.}$$

The test statistic is

$$F_{H_{05}} = \frac{SSH_{05}/(t - 1)}{SSRes/(n - 2t)}. \tag{2.27}$$

The conclusion one makes at X = 0 may be different from that made at

$$X = \bar{x}.. \text{ or } X = X_0,$$

where X_0 is some other fixed value of the covariate. Suppose the experimenter wants to compare the distances between the lines at a selected value of X, say $X = X_0$. The hypothesis to be tested is

$$H_{06}\text{: } \mu_{Y_1|\beta_i X = \beta_i X_0} = \mu_{Y_2|\beta_i X = \beta_i X_0} = \cdots \mu_{Y_t|\beta_i X = \beta_i X_0} = \mu_{X_0}\ , \tag{2.28}$$

where

$$\mu_{Y_i|\beta_i X = \beta_i X_0} = \alpha_i + \beta_i X_0 .$$

The model in 2.1 can be equivalently expressed as

$$\begin{aligned} y_{ij} &= \alpha_i + \beta_i X_0 - \beta_i X_0 + \beta_i X_{ij} + \varepsilon_{ij} = \mu_{Y_i|\beta_i X = \beta_i X_0} \\ &+ \beta_i (X_{ij} - X_0) + \varepsilon_{ij}. \end{aligned} \tag{2.29}$$

The model restricted by H_{06} is

$$y_{ij} = \mu_{Y|\beta_i X = \beta_i X_0} + \beta_i(X_{ij} - X_0) + \varepsilon_{ij}' \tag{2.30}$$

and the corresponding residual sum of squares is

$$SSRes(H_{06}) = \sum_{i=1}^{t} \sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_{X_0} - \beta_i(X_{ij} - X_0))^2, \tag{2.31}$$

which is based on

$$d.f._{SSRes(H_{06})} - d.f._{SSRes} = N - t - 1 \text{ degrees of freedom.}$$

The sum of squares due to deviations from H_{06} is $SSH_{06} = SSRes(H_{06}) - SSRes$, which is based on

$$d.f._{SSRes(H_{06})} - d.f._{SSRes} = t - 1 \text{ degrees of freedom.}$$

The resulting test statistic is

$$F_{H_{06}} = \frac{(SH_{06})t - 1}{SSRes/(N - 2t)}. \tag{2.32}$$

It is important for the experimenter to make comparisons between the lines at several different values of the covariate. The usual comparison of adjusted means, i.e., at

$$X = \bar{x}..$$

is only one of many comparisons which are probably of interest. Figure 2.3 shows three possible values of X (covariate) at which one might make comparisons. The corresponding test statistics can be obtained by expressing

$$H_{06} \text{ with } X_0 = X_t, X_0 = \bar{x}.. \text{ and } X_0 = X^*$$

respectively. If a hypothesis corresponding to a selected value of X_0 is rejected, then a multiple comparison procedure could be used to determine which distances between pairs of lines are not equal to zero or the preplanned treatment comparisons could be made at X_0. The standard error of the difference between two lines at

$$X = X_0, \hat{\mu}_{y_i|\beta_i X = \beta_i X_0} - \hat{\mu}_{Y_{i'}|\beta_i X = \beta_i X_0},$$

is

$$S_{\hat{\mu}_{Y_i|\beta_i X = \beta_i X_0} - \hat{\mu}_{Y_{i'}|\beta_{i'} X = \beta_{i'} X_0}} = \left[S^2_{\hat{\mu}_{Y_i|\beta_i X = \beta_i X_0}} + S^2_{\hat{\mu}_{Y_{i'}|\beta_{i'} X = \beta_{i'} X_0}}\right]^{1/2} \tag{2.33}$$

where the standard errors $S_{\hat{\mu}_{Y_i|\beta_i X = \beta_i X_0}}$ can be obtained from the standard errors of the intercept parameters in model (2.29). These standard errors are computed with the assumption the two models have no common parameters so the covariance between the two adjusted means is zero. Again, preplanned comparisons and LSD, Bonferroni or Scheffe types of multiple comparison procedures can be used to help interpret the results of the analysis of covariance. For most experiments, comparisons should be made for at least three values of X, one in the lower range, one in the middle range and one in the upper ranges of the X's obtained for the experiment.

The experimenter is often interested in determining which treatment mean responds the most to a change in the value of the covariate. In this case an LSD approach can be used to make size α comparison wise tests about the β_i's. Alternatively, one could also use a Bonferroni or Scheffe-type approach to control the experiment wise error rate. In any case, the standard error of the difference between two slopes is

$$S_{\beta_i - \beta_i} = \left[S^2_{\beta_i} + S^2_{\beta'}\right]^{1/2}$$

where S_{β_i} denotes the standard error associated with $\hat{\beta}_i$ and it is assumed the covariance between the two parameters is zero (which is the case here since the two models do not have common parameters.) The degrees of freedom for the appropriate percentage point is N - 2t. Preplanned treatment comparisons can be made by comparing the slopes of the various models. Section 2.7 shows how to carry out the computations via SAS.

2.5 Confidence Bands About the Difference of Two Treatments

When the slopes are unequal, it is often useful to determine the region of the covariate where the two treatments produce significantly different responses. A confidence band can be constructed about the difference of two treatment models and the region of the covariate where the confidence band does not contain zero is determined. A Scheffe type confidence statement should be used to provide experimentwise error protection. The difference between two lines for treatment 1 and 2 at X = X_0 is

$$\hat{\mu}_{Y_1|\beta_1 X = \beta_1 X_0} - \hat{\mu}_{Y_2|\beta_2 X = \beta_2 X_0} = \hat{\alpha}_1 + \hat{\beta}_1 X_0 - \hat{\alpha}_2 - \hat{\beta}_2 X_0$$

which has standard error

$$S_{1-2|\beta_i X_0} = \left[S^2_{\hat{\mu}_{Y_1|\beta_1 X = \beta_1 X_0}} + S^2_{\hat{\mu}_{Y_2|\beta_2 X = \beta_2 X_0}}\right]^{1/2}$$

where

$$S^2_{\mu_{Y_i|\beta_i X = \beta_i X_0}} = S^2_{\hat{\alpha}_i} + X_0^2 S^2_{\beta_i} + 2X_0 \mathrm{Cov}(\hat{\alpha}_i, \hat{\beta}_i).$$

An example of the construction and use of the confidence band about the difference of two regression lines is in Section 2.8.

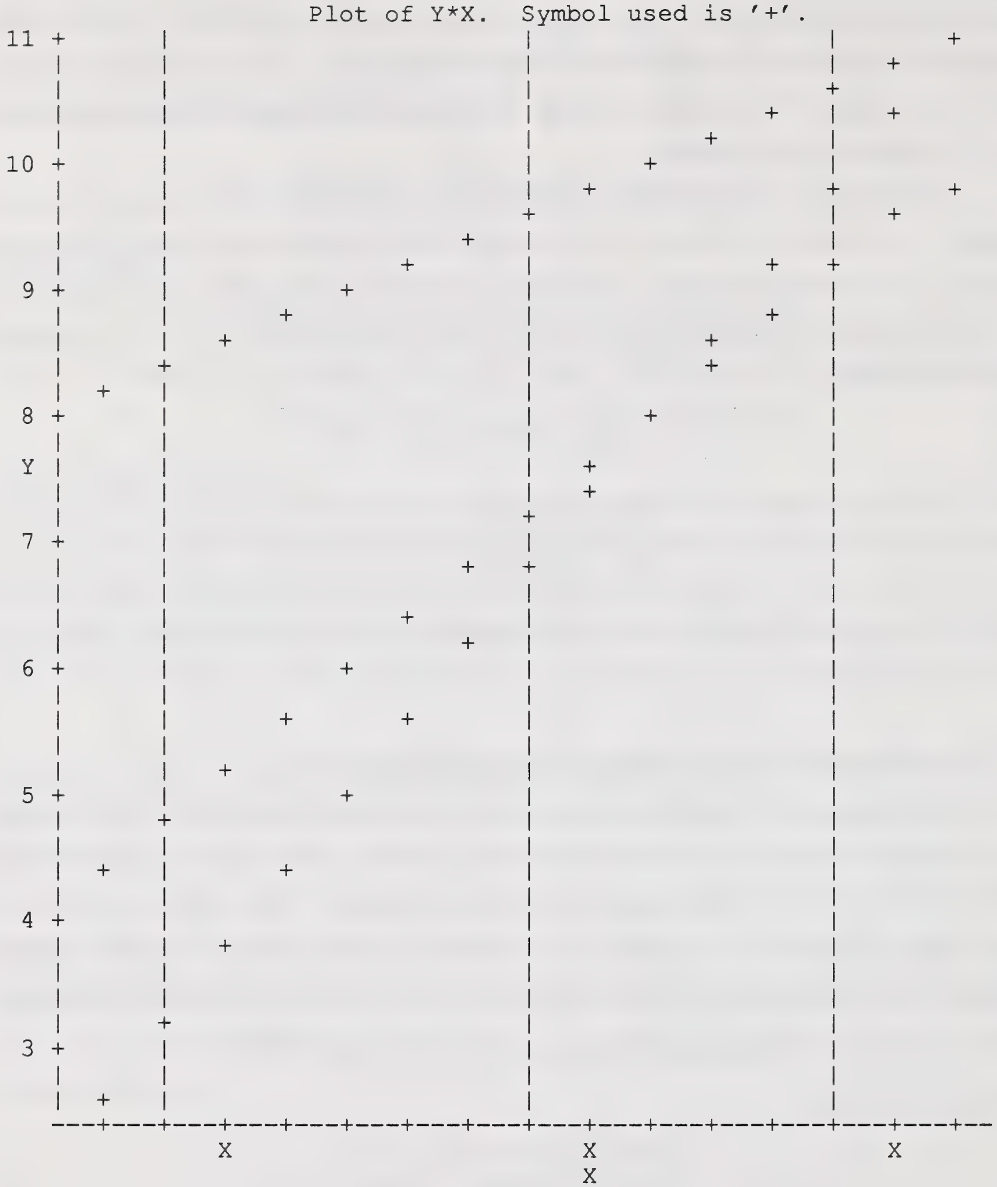


Figure 2.3 For non-parallel lines models the comparisons between the treatments depends of the selected value of X.

2.6 Summary of Strategies

Sections 2.3 and 2.4 describe the strategies for determining the form of the model and the resulting analyses. Table 2.6.1 lists the paths for model determination, but there are possible exceptions.

Table 2.6.1 Strategy for determining the form of the analysis of covariance model involving one covariate assuming a simple linear regression model will describe each treatment data.

a) Test the hypothesis that the slopes are zero
- i) If fail to reject, compare the treatments via analysis of variance,
- ii) if reject go to b)

b) Test the hypothesis that the slopes are equal
- i) if fail to reject, use a parallel lines model and compare the treatments by comparing the intercepts or adjusted means (LSMEANS).
- ii) if reject to c)

c) Use the unequal slope model and
- i) compare the slopes of the treatments to see if treatments can be grouped into groups with equal slopes
- ii) compare the models of at least three values of the covariate, low, middle and high value
- iii) construct confidence bands about the difference of selected pairs of models.

There are at least two possible exceptions to the strategy. First, it is possible to reject the equal slope hypothesis and fail to reject the zero slope hypothesis when there are both positive and negative slopes. In this case, use the non-parallel lines model. Second, it is possible to fail to reject the zero slopes hypothesis when in fact the slope of the parallel lines model is significantly different from zero. Here there is not enough information from the individual treatments to say the slopes are different from zero, but the combined information does detect the linear relationship. In this case use the common slope or parallel lines model.

2.7 Analysis of Covariance Computations via the SAS System

The SAS System can be used to compute the various estimators and tests of hypotheses discussed in the previous sections. The SAS System statements required for each part of the analysis are presented in this section and detailed examples are discussed in Sections 2.8 and 2.9.

All the following models will be fit assuming that the data were read in by the following statements:

```
DATA ONECOV;INPUT TRT  Y  X;
CARDS;
```

The required SAS System statements needed to fit model (2.1) are

```
PROC GLM; CLASSES TRT;
MODEL Y = TRT X*TRT /NOINT SOLUTION;
```

The term TRT with the no-intercept (NOINT) option generates the part of the design matrix corresponding to the intercepts and enables one to obtain the estimators of the intercepts. The term X*TRT generates the part of the design matrix corresponding to the slopes. The SOLUTION option is used so that the estimators and their standard errors are printed. (PROC GLM does not provide the estimators when there is a CLASS variable unless SOLUTION is specified.) The sum of squares corresponding to ERROR is SSRes of equation (2.4) and the MEAN SQUARE ERROR is $\hat{\sigma}^2$, the estimate of the sampling variance.

The type III (or Type IV) sum of squares corresponding to X*TRT tests H_{01} of (2.6). The Type III (or Type IV) sum of squares corresponding to TRT tests H_{05}, i.e., $\alpha_1 = \alpha_2 = ... = \alpha_t = 0$ given that the unequal slopes are in the model. This hypothesis is often not of interest, but a test is available in case the experimenter has a situation where a zero intercept hypothesis is interpretable.

Next, to test H_{02} of (2.11), the required SAS System statements are:

```
PROC GLM; CLASSES TRT;
MODEL Y = TRT X X*TRT / SOLUTION;
```

The type III sum of squares corresponding to X*TRT tests H_{02}. The type III sum of squares corresponding to X tests if the average value of the slopes is zero and the type III sums of squares corresponding to TRT tests H_{05}. By including X and/or removing the NOINT option, the model is singular and the provided least squares solutions is not directly interpretable. The obtained least squares solution satisfies the set-to-zero restrictions (see chapter 6 of Milliken and Johnson (1984)). If one uses the model statement Y = TRT X*TRT X, where X*TRT is listed before X, the type I sum of squares corresponding to X*TRT tests H_{01} while the type III sum of squares tests H_{02}. A list of type I and III estimable functions can be obtained and used to verify the hypothesis tested by each sum of squares. If one fails to reject H_{02}, the parallel lines or equal slope model of 2.12 should be fit to the data. The appropriate SAS System statements are:

```
PROC GLM; CLASS TRT;
MODEL Y = TRT X /SOLUTION;.
```

The type III sum of squares corresponding to TRT is SSH_{03} of (2.21), and the resulting F ratio tests that the distances between the lines are zero given that the parallel line model is adequate to describe the data.

Estimates of the mean of y given $X = \bar{x}_{..}$ for each treatment, which are often called adjusted means, can be obtained by including the statement

```
LSMEANS TRT / STDERR PDIFF;
```

after the MODEL statement. This provides the adjusted means

$$\hat{\mu}_{Y_i|\beta_i X = \beta_i \bar{x}} = \hat{\alpha}_i + \hat{\beta}_i \bar{x}$$

the estimate of the treatment mean at X = $\bar{x}$ or least squares mean, for each treatment and the option STDERR provides the corresponding standard errors of the adjusted means. The PDIFF option provides significance levels for t-tests of

$$\mu_{Y_i|\beta_i X = \beta_i \bar{x}..} = \mu_{Y_r|\beta_r X = \beta_i \bar{x}} \cdot$$

(The TDIFF option provides the values of the t-tests from which the PDIFF values are obtained.) for each pair of adjusted means. A comparison of adjusted means is also a comparison of the α_i's for the parallel lines model. The significance probabilities can be used to construct a LSD or a Bonferroni multiple comparison procedure for comparing the distances between pairs of lines.

Any comparisons between parameters can be made by using the ESTIMATE or CONTRAST statement in the GLM procedure. There are two situations where such statements are needed.
First, if the conclusion is that the slopes are not equal, then one can apply a multiple comparison procedure in order to compare some or all pairs of slopes. This is easily done by including an ESTIMATE statement following the MODEL statement for each comparison of interest. For example, if there are three treatments and it is of interest to compare all pairs of slopes, then the following statements would be used:

```
PROC GLM; CLASSES TRT;
MODEL Y=TRT X*TRT / NOINT SOLUTION;
ESTIMATE 'B1-B2' X*TRT  1  -1  0;
ESTIMATE 'B1-B3' X*TRT  1   0 -1;
ESTIMATE 'B2-B3' X*TRT  0   1 -1;
```

Each ESTIMATE statement produces an estimate of the linear combination of parameters, a computed t-value, and it's significance level using the residual mean square as the estimate of σ^2. The significance levels obtained from these comparisons can be used to construct a LSD or Bonferroni multiple comparison procedure.

For the unequal slope model, adjusted means needs to be obtained X = x.. and at other specified values of X. Such adjusted means can be used to make comparisons between the treatments. The SAS system code for using ESTIMATE statements to obtain adjusted means at X = 7.3 is

```
ESTIMATE  'T1 AT 7.3'  TRT 1 0 0 X*TRT 7.3  0    0;
ESTIMATE  'T3 AT 7.3'  TRT 0 1 0 X*TRT  0  7.3   0;
ESTIMATE  'T3 AT 7.3'  TRT 0 0 1 X*TRT  0    0  7.3;.
```

Contrasts between the adjusted means at X = 7.3 can be obtained by subtracting the respective estimate statement values from the adjusted mean estimate statements as

```
ESTIMATE  'T1 - T2 AT 7.3'  TRT 1 -1 - X*TRT 7.3  -7.3  0;
ESTIMATE  'T1 - 2T + T3 AT 7.3'  TRT 1 -2 1 X*TRT 7.3 -14.6 7.3;.
```

The SAS system can be used to construct confidence bands about the difference of two models. Since it is unlikely that the values of the covariates will provide a uniform coverage of the covariate region, one must add to the data set additional observations for each treatment corresponding to the desired values of the covariate (y is assigned a missing value for these observations). The following SAS system code generates one set of values to be added to the data set; fits three models and constructs the confidence band about the difference between the models of treatments 1 and 2.

```
DATA RAW; INPUT  TRT  y x;
CARDS
(THE DATA)
DATA GENERATE;  Y = .;
        DO TRT = 1 TO 3;
                DO X = 1 TO 10;
                OUTPUT;
                END;
        END;
DATA ALL; SET RAW GENERATE;
PROC GLM; CLASS TRT;
MODEL Y = TRT X*TRT/SOLUTION;
OUTPUT OUT = VALUE P = Y STDP = STD;
DATA ONE; SET VALUE; IF TRT = 1;
PY1 = PY; STD1 = STD;
PROC SORT; BY X;
DATA TWO; SET VALUE; IF TRT = 2;
PY2 = PY; STD2 = STD;
PROC SORT; BY X;
DATA ONETWO; MERGE ONE TWO; BY X;
Q = SQRT(2*INVF(2,27));
DIF = PY1 - PY2;
STEDIF = SQRT(STD1**2 + STD2**2);
LOW = DIFF - Q*STEDIF;
HIGH = DIFF + Q*STEDIF;
PROC PLOT; PLOT
        DIFF*X = '*' LOW*X = '-' HIGH*X = '-'
       /OVERLAY;
```

The value of Q in the SAS system code corresponds to the Scheffe percentage point. The analysis and the SAS System computations described in this section are demonstrated in detail by examples in the next sections.

2.8 Example: Sugar Beets and Nitrogen--Equal Slopes

In a study of growth regulators for sugar beets, it was determined that there was substantial plot to plot variation in the level of available nitrogen. The amount of nitrogen in the soil can affect the yield of the beets in addition to an effect due to the growth regulators. After planting the sugar beets, the available nitrogen was measured from soil samples obtained from each plot. Since there is a lot of plot to plot variation in the amount of available nitrogen, the experimenter wanted to be able to compare the effect of the growth regulators at specified levels of nitrogen. Thus, the available nitrogen in the soil is used as the covariate and the dependent variable is the yield of the sugarbeet roots per plot in pounds. The experimental design is a one-way treatment structure with three growth regulators (TRT) and 10 plots per treatment in a completely randomized design structure. The data are in Table 2.8.1.

Table 2.8.1. Plot Yield and Residual Nitrogen for the Sugar Beet Example

TRT	1		2		3	
	YIELD	NITROGEN	YIELD	NITROGEN	YIELD	NITROGEN
	210	100	155	65	155	55
	150	50	150	55	160	55
	200	105	170	65	165	65
	180	75	175	75	185	70
	190	80	185	75	185	80
	220	110	195	90	200	85
	170	60	205	84	205	90
	170	70	210	100	215	105
	190	90	215	95	220	105
	220	100	220	100	220	100

The SAS system statements required to carry out the analysis are given in Table 2.8.2 along with an annotation of each statement's function. The analysis of variance tables obtained and the estimates of the parameters are given in Tables 2.8.3 to 2.8.5.

Table 2.8.2. SAS System Statements for Sugar Beet - Nitrogen Example

i) Read in the data

```
DATA COV; INPUT TRT YIELD NIT;
CARDS;
```

ii) Fit the model to test H in (2.6)

```
PROC GLM; CLASSES TRT;
MODEL YIELD=TRT NIT*TRT / NOINT SOLUTION;
```

iii) Fit model to test H in (2.11)

```
PROC GLM; CLASSES TRT;
MODEL YIELD=TRT NIT NIT*TRT;
```

iv) Fit parallel lines model to compare adjusted means

```
PROC GLM; CLASSES TRT;
MODEL YIELD=TRT NIT / SOLUTION;
LSMEANS TRT / STDERR PDIFF;
```

Table 2.8.3 Results from Fitting Model [2.1] to Data for Example 1 for Parameter Estimation and to Test H_{01} of (2.6). From Part (ii) of Table 2.8.2.

DEPENDENT VARIABLE: YIELD

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	6	1094097.45	182349.56
ERROR	24	1101.55	45.64

SOURCE	DF	TYPEIII SS	F VALUE	PR>F
TRT	3	1079250.00	7830.94	.0001
NIT*TRT	3	14847.45	107.73	.0001

PARAMETER		ESTIMATE	STD ERROR OF EST
TRT	1	98.943	9.61
	2	66.320	11.64
	3	88.780	9.74
NIT*TRT	1	1.084	0.11
	2	1.512	0.14
	3	1.262	0.12

Table 2.8.4. Results of Fitting Model (2.1) to Data for Sugar Beet - Nitrogen Example for Parallel Line Analysis

(MODEL and ERROR sums of squares and mean squares are the same as those given in Table 2.8.3.)

SOURCE	d.f.	TYPE IV SS	F VALUE	PR>F
TRT	2	217.278	2.36	0.1155
NIT	1	14726.736	320.57	0.0001
NIT*TRT	2	257.555	2.80	0.0805

Table 2.8.5 Results from Fitting Model (2.13) to Data for Sugar Beet - Nitrogen Example for Parallel Line Analysis

DEPENDENT VARIABLE: YIELD

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	3	14636.56	4878.85
ERROR	26	1360.10	52.31
CORRECTED TOTAL	29	15996.66	

	DF	TYPEIV SS	F VALUE	PR>F
SOURCE	2	112.60	1.08	0.3556
TRT	1	14589.89	278.90	0.0001

PARAMETER		ESTIMATE	T FOR H0: PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
INTERCEPT		89.56 B	13.80	0.0001	6.49
TRT	1	-4.76 B	-1.47	0.1543	3.24
TRT	2	-2.37 B	-0.73	0.4696	3.23
TRT	3	0.00 B	.	.	.
NIT		1.25	16.70	0.0001	0.07

TRT	OBSERVED MEAN	ADJUSTED MEAN	STDERR(ADJ MN)	PROB>|T| H0:LSMEAN=0
1	190.0	187.28	2.293	0.0001
2	188.0	189.67	2.289	0.0001
3	191.0	192.04	2.288	0.0001

t-tests for differences between adjusted mean
PROB > |T| H0: LSMEAN(I) = LSMEAN(J)

I/J	1	2	3
1	.	0.4693	0.1543
2	0.4693	.	0.4696
3	0.1543	0.4696	.

Model (2.1) was fit to the data with the results appearing in Table 2.8.3. The sums of squares for NIT*TRT tests H_{01} of (2.6) and is highly significant. Thus the experimenter would conclude that the expected mean yields do depend on NIT (available nitrogen). Table 2.8.3 also displays the estimates of the model parameters, $\hat{\alpha}_i$'s correspond to TRT and $\hat{\beta}_i$'s correspond to NIT*TRT.

Next, the equality of slopes hypothesis H_{02} of (2.11) is tested to determine if a parallel lines model will adequately describe the data. The sum of squares due to NIT*TRT in Table 2.8.4 is SSH_{02}. The significance level of the test statistic is .0805. If the decision is made at the .05 level, then one would conclude that a parallel lines model is adequate to describe the data. However, if the decision is made at the .10 level, then one would conclude that the parallel lines model is not adequate. For the purpose of discussion, the .05 level is used and the parallel lines model is determined to be adequate to describe the relationship between yield and level of available nitrogen (one should plot the residuals before continuing).

The parallel lines model of (2.13) was fit to the data and the results are given in Table 2.8.5.

The sum of squares due to TRT is testing H_{03} of (2.17) given that H_{02} is true. The significance level is .3556, indicating that the distances between the regression lines are not significantly different from zero. The estimated slope of the parallel regression lines is 1.252, which is significantly different from zero.

Table 2.8.5 also contains the observed TRT means for yield and the adjusted means of the TRT yields. The adjusted means, provided by the LSMEANS statement, are the predicted values obtained from the treatment regression lines evaluated at the mean NIT (nitrogen) value for the experiment. The mean NIT value for the experiment is 81.83. For example, the adjusted mean for TRT 1 is

$$\hat{\mu}_{Y_1|NIT = 81.833} = (89.56 - 4.76) + 81.83(1.25) = 187.65.$$

The quantity (89.56 - 4.76) = 84.80 is the intercept ($\hat{\alpha}_1$) for treatment 1 from the parallel lines model.

The significance levels of the t-tests for comparing each pair of adjusted treatment means is presented in Table 2.8.5. The significance levels are obtained by including the PDIFF option in the LSMEANS statement. The t-statistics to compare adjusted means for treatments i and i' are computed as

$$t_c = \frac{\text{Difference between two adjusted means}}{\text{Standard error of the difference of two adjusted means}}$$

$$= \frac{\hat{\mu}_{Y_i|NIT = 81.833} - \hat{\mu}_{Y_{i'}|NIT = 81.833}}{\hat{S}_{\hat{\mu}_{Y_i}|NIT = 81.833 - \hat{\mu}_{Y_{i'}}|NIT = 81.83333}}$$

where the standard error is computed via (2.23).

The probability values can be used to construct an LSD procedure by concluding the distance between two lines is significantly different from zero if the printed significance level is less than, say, .05. A Bonferroni multiple comparison procedure can be used by concluding the distance between two lines is significantly different from zero if the printed significance level is less than .05/3 = .01667. Of course, the above decisions could be made at some significance level other than .05.

For this data set, none of the lines are significantly different from each other, therefore, one line could be used to describe all of the data. Thus, conclude that there are no differences between the growth regulator treatments. A graph of the data and estimated regression lines are in Figure 2.4.

2.9 Example--Exercise Programs and Initial Resting Heart Rate

An exercise physiologist structured three types of exercise programs (EPRO) and conducted an experiment to evaluate and compare the effectiveness of each program. The experiment consisted of subjecting a person to a given exercise program for eight weeks. At the end of the training program, each person ran for six minutes after which their heart rate was measured. An exercise program is determined to be more effective if individual's on that program have lower heart rates than individuals on another exercise program. Since people entered the experiment at differing degrees of fitness, the resting heart rate before beginning training was recorded, and was used as a covariate. The object of the study is to be able to compare exercise programs at a common initial resting heart rate. To carry out the experiment, 24 males between 28 and 35 years of age were selected and then 8 males were randomly assigned to each of the three EPRO treatments. The exercise program (EPRO), heart rate (HR) after the 6 minutes run at the completion of eight weeks of training and the initial resting heart rate (IHR) are given in Table 2.9.1.

Table 2.9.1 Data for Exercise Program Example

EXERCISE PROGRAM					
1		2		3	
HR	IHR	HR	IHR	HR	IHR
118	56	148	60	153	56
138	59	159	62	150	58
142	62	162	65	158	61
147	68	157	66	152	64
160	71	169	73	160	72
166	76	164	75	154	75
165	83	179	84	155	82
171	87	177	88	164	86

The SAS system statements necessary to analyze this data set are given in Table 2.9.2. Model 2.1 was first fit to the data and the results appear in Table 2.9.3. The sum of squares for the IHR*EPRO tests H_{01} of (2.6) and is highly significant. Thus, conclude that the heart rate depends on the initial resting hear rate. Table 2.9.3 also contains the estimates of the parameters,

$\hat{\alpha}_i$'s corresponding to EPRO and $\hat{\beta}_i$'s corresponding to IHR*EPRO.

Table 2.9.2 SAS System Statements for Exercise Program Example

i) Read in the data cards.

```
DATA COVI; INPUT EPRO HR IHR; CARDS;
```

ii) Fit model to test H_{01} in (2.6) and to provide estimate of several linear combinations of the parameters

```
PROC GLM: CLASSES EPRO;
MODEL HR=EPRO IHR*EPRO/NOINT SOLUTION
LSMEANS EPRO/STDERR PDIFF;
ESTIMATE ’B1-B2, IHR*EPRO 1 -1 0;
ESTIMATE ’B1-B3’ IHR*EPRO 1 0 -1;
ESTIMATE ’B2-B3’ IHR*EPRO 0 1 -1:
ESTIMATE ’1 AT 60 - 2 AT 80; EPRO 1 -1 0 IHR*EPRO 60 -80 0:
ESTIMATE ’A1-A2’ EPRO 1 -1 0;
ESTIMATE ’A2-A3’ EPRO 0 1 -1;
ESTIMATE ’A1-A3’ EPRO 1 0 -1;
ESTIMATE ’1-2 AT 55’ EPRO 1 -1 0 IHR*EPRO 55 -55 0;
ESTIMATE ’1-2 AT 70’ EPRO 1 -1 0 IHR*EPRO 70 -70 0;
ESTIMATE ’1-2 AT 85’ EPRO 1 -1 0 IHR*EPRO 85 -85 0;
ESTIMATE ’1-3 AT 55’ EPRO 1 0 -1 IHR*EPRO 55 0 -55;
ESTIMATE ’1-3 AT 70’ EPRO 1 0 -1 IHR*EPRO 70 0 -70;
ESTIMATE ’1-3 AT 85’ EPRO 1 0 -1 IHR*EPRO 85 0 -85;
ESTIMATE ’2-3 AT 55’ EPRO 0 1 -1 IHR*EPRO 0 55 -55;
ESTIMATE ’2-3 AT 70’ EPRO 0 1 -1 IHR*EPRO 0 70 -70;
ESTIMATE ’2-3 AT 85’ EPRO 0 1 -1 IHR*EPRO 0 85 -85;
```

iii) Fit model to test H_{02} in (2.11)

```
PROC GLM; CLASSES EPRO;
MODEL HR = EPRO IHR  IHR*EPRO / SOLUTION NOINT;
```

Table 2.9.3 **Results From Fitting Model (2.1) to the Data in the Exercise Program Example for Parameter Estimation and to Test H_{01} of (2.6) [Part (ii) of Table 2.9.2]**

DEPENDENT VARIABLE: HR

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	6	594974.35	99162.38
ERROR	18	507.64	28.20
UNCORRECTED ERROR	24	595482.00	

SOURCE	DF	TYPE IV SS	F VALUE	PR>F
EPRO	3	5108.68	60.38	0.0001
IHR*EPRO	3	2650.60	31.33	0.0001

PARAMETER		ESTIMATE	T FOR H0: PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
EPRO	1	46.53	3.66	0.0018	12.71
	2	97.19	6.88	0.0001	14.12
	3	137.48	10.97	0.0001	12.52
IHR*EPRO	1	1.48	8.29	0.0001	0.17
	2	0.93	4.80	0.0001	0.19
	3	0.26	1.47	0.1576	0.17

Next the experimenter needs to determine if a parallel lines model will adequately describe the data, i.e., test H_{02} of (2.11). The sum of squares due to IHR*EPRO in Table 2.9.4 is SSH_{02} (from part iii of Table 2.9.2). The significance level of the test is .0006, indicating that the parallel lines model is not reasonable to describe this data. Thus, it is necessary to compare the EPRO at various values of IHR by using the unequal slope model. The experimenter chose to compare the three exercise programs at IHR = 55, 70 and 85. In Table 2.9.2, the last 9 ESTIMATE statements are those required to make the three pairwise comparisons between the EPR's at each given value of IHR. For example, '1-2 AT 55' asks GLM to compare EPRO 1 with EPRO 2 at IHR = 55, etc. The results of these nine tests are in Table 2.9.5. Table 2.9.5 also contains the results of testing the following special hypotheses:

i) $\beta_1 = \beta_{2}$· $\beta_1 = \beta_3$ and $\beta_2 = \beta_3$

ii) $\mu_{Y_1|x=60} - \mu_{Y_2|x=80}$

iii) $\alpha_1 = \alpha_{2}$·, $\alpha_1 = \alpha_3$ and $\alpha_2 = \alpha_3$

Investigating the relationship between the slopes is reasonable as one might wish to simplify the model by grouping treatments with "equal" slopes into groups where a parallel lines model is fit to the data within a group and nonparallel lines between groups. When such a simplification can occur, the comparison process is easier as one can compare lines within a group with the LSMEANS.

Comparing two treatments at different values of the covariate (ii from above) is not usually done, but there are some circumstances where such comparisons are warranted (like comparing EPRO 1 at 60 to EPRO 2 at 80, not meaningful here). But, for example, suppose it costs the same for X_1 units of the covariate on process 1 as for X_2 units of the covariate on process 2, then it is reasonable to compare the models at different values of X, but on an equal cost basis. Comparison of the intercepts (part iii from above) compares the distances between the regression lines at IHR = 0. In this example, it is stupid to compare the intercepts. We must be careful not be make such stupid comparisons.

Table 2.9.4 Results of Fitting Model (2.1) to Exercise Program Example to Test H_{02} of (2.6)

(Model and ERROR sum of squares and the mean squares are the same as Table 2.8.

SOURCE	DF	TYPE IV SS	F VALUE	PR>F
EPRO	3	3108.68	60.38	0.0001
IHR	1	1990.74	70.59	0.0001
IHR*EPRO	1	658.98	11.68	0.0006

The adjusted means in Table 2.9.5 estimate the EPRO means at IHR = 70.375 the overall mean IHR for the experiment. While it might be of interest to compare the treatment means at 70.375, they should also be compared to other values of the covariate as well. Figure 2.5 summarizes the analysis with the estimated regression lines and the comparisons of means at IHR = 55, 70 and 85. If two means at the same value of IHR are followed by the same letter, the means are not significantly different while unlike letters indicate a difference. A .05 Bonferroni approach was used at each level of IHR where a difference is declared significant within a IHR value if the t-test has a significance level less that or equal to .05/3 = .01667.

Another way to compare the regression lines is to construct confidence bands about the difference of each pair of models and determine the values of the independent variable where the bands exclude zero. To provide control over the error rates, Scheffe percentage points are used in constructing the bands. The difference between two models (ith and i'th) is $(\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)X$ which depends on two parameters $(\alpha_1 - \alpha_2)$ and $(\beta_1 - \beta_2)$. Thus the percentage point used is

$$(2\ F_{\alpha,2,\upsilon})^{1/2}$$

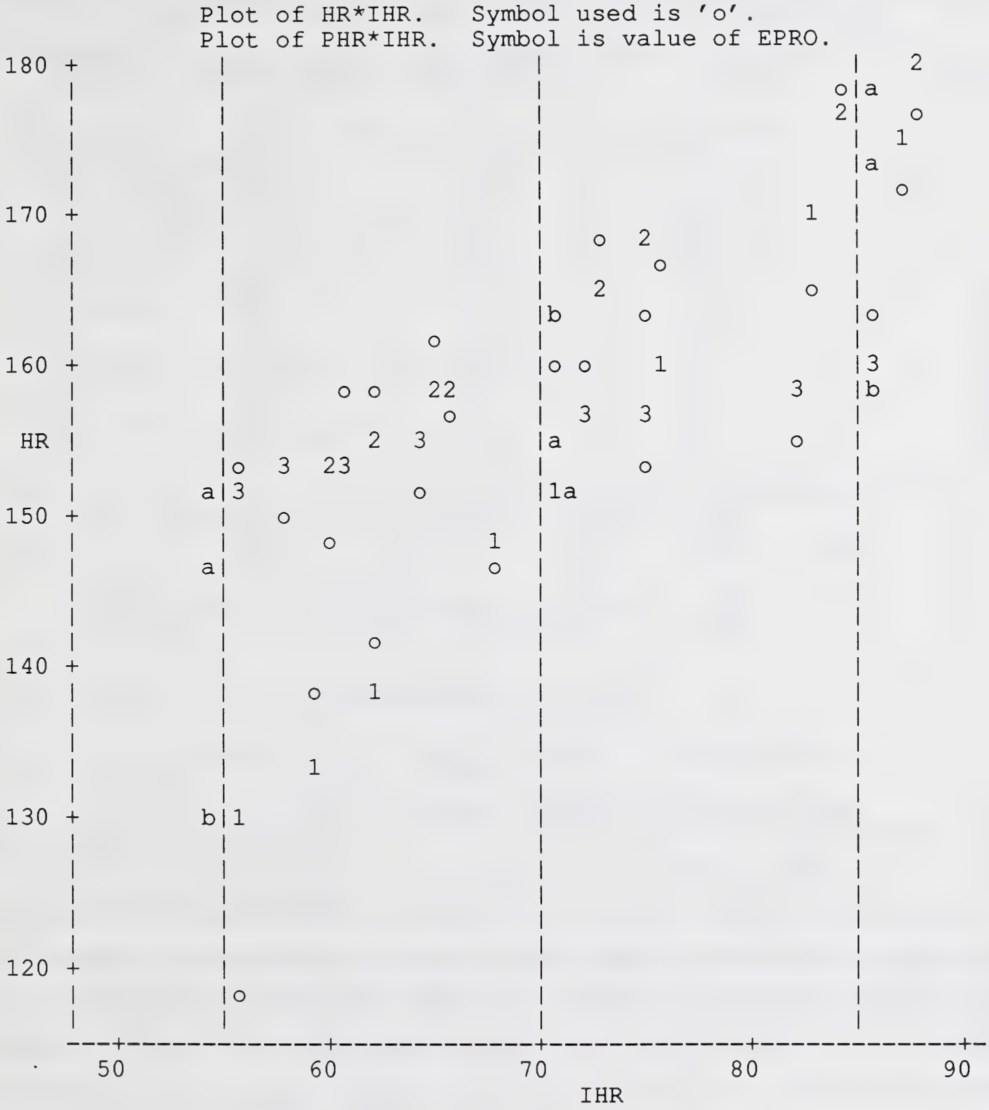


Figure 2.5 Graphs of the estimated models for the exercise program example. Bonferroni comparisons between the exercise programs are shown for IHR = 55, 70, 85, where models with the same letter are not significantly different at α=.05. Comparisons are obtained from Table 2.9.5.

Table 2.9.5 **Estimates of Selected Linear Combinations of the Parameters for Exercise Program Example**

PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
B1-B2	0.54	2.06	0.0537	0.26
B1-B3	1.22	4.83	0.0001	0.25
B2-B3	0.67	2.54	0.0203	0.26
1 AT 60 - 2 AT 80	-36.57	-10.11	0.0001	3.61
A1-A2	-50.65	-2.66	0.0158	19.01
A2-A3	-40.29	-2.13	0.0469	18.88
A1-A3	-90.94	-5.09	0.0001	17.85
1-2 AT 55	-20.55	-4.11	0.0007	5.00
1-2 AT 70	-12.34	-4.62	0.0002	2.67
1-2 AT 85	-4.13	-0.91	0.3769	4.56
1-3 AT 55	-23.76	-5.19	0.0001	4.58
1-3 AT 70	-5.44	-2.05	0.0555	2.65
1-3 AT 85	12.87	2.75	0.0132	4.68
2-3 AT 55	-3.21	-0.65	0.5215	4.81
2-3 AT 70	6.90	2.58	0.0190	2.67
2-3 AT 85	17.01	3.64	0.0019	4.67

EPRO	HR LSMEAN	STD ERR LSMEAN	PROB>|T| H0:LSMEAN=0	LSMEAN NUMBER
1	151.06	1.87	0.0001	1
2	163.20	1.89	0.0001	2
3	156.04	1.88	0.0001	3

PROB > |T| H0: LSMEAN(I)=LSMEAN(J)

I/J	1	2	3
1	.	0.0002	0.0775
2	0.0002	.	0.0154
3	0.0775	0.0154	.

where υ is the degrees of freedom associated with the MSERROR. Since the estimates of the parameters for the models are independent between the models, the variance of the difference of two models is the sum of the variances of the two predicted values. Most computer codes provide the predicted values and the standard errors of the predicted values which can be combined to construct the confidence band about the difference of two models. The end points of the confidence band at a given value of the independent variable is

$$(\hat{\alpha}_1+\hat{\beta}_1X) - (\hat{\alpha}_2+\hat{\beta}_2X) \pm (2\ F_{\alpha/2,2,\upsilon})^{1/2}(\mathrm{var}(\hat{\alpha}_1+\hat{\beta}_1X) + \mathrm{var}(\hat{\alpha}_2+\hat{\beta}_2X))^{1/2}.$$

The annotated SAS system code used to generate the data sets at values of IHR from 55 to 85 (and generate the necessary missing values for HR) and construct the confidence bands is displayed in Table 2.9.6. The three graphs for comparing all pairs of models are in Figures 2.6 to 2.8.

Table 2.9.6 SAS System code for constructing confidence bands about the difference of pairs of treatment (EPRO) models.

a) Generate a data set with IHR values from 55 to 80 for each EPRO.

```
DATA PLOT; HR=.;
DO EPRO= 1 TO 3;
DO IHR=55 TO 85;
OUTPUT;
END;
END;
```

b) Merge the generated data with the actual data set.

```
DATA ALL; SET HEART PLOT;
```

c) Use PROC GLM to estimate the parameters of the models and compute the predicted values and their standard errors. Write the results to a data set -- XPLOT.

```
PROC GLM; CLASSES EPRO;
MODEL HR=EPRO IHR*EPRO/NOINT SOLUTION;
OUTPUT OUT=XPLOT P=PHR STDP=STDR;
```

d) For data sets for each treatments data.

```
DATA ONE; SET XPLOT; IF EPRO=1 AND HR=.;P1=PHR;S1=STDR;
PROC SORT; BY IHR;
DATA TWO; SET XPLOT; IF EPRO=2 AND HR=.;P2=PHR;S2=STDR;
PROC SORT; BY IHR;
DATA THR; SET XPLOT; IF EPRO=3 AND HR=.;P3=PHR;S3=STDR;
PROC SORT; BY IHR;
```

e) Merge pairs of data sets by the IHR, compute the difference between the models at each IHR, compute the standard error of each difference, compute the upper and lower limits, and plot the results. The value 2.665 =$(2 * 3.55)^{\frac{1}{2}}$, the Scheffe percentage point for each pair of models. The SAS system code is included for comparing models 1 and 2.

```
DATA ONE_TWO; MERGE ONE TWO; BY IHR;
DIFF=P1-P2;STE=SQRT(S1*S1+S2*S2); Z=0;
LOW=DIFF-2.665*STE;
HIGH=DIFF+2.66*STE;
PROC PLOT; PLOT DIFF*IHR=’*’ LOW*IHR=’+’ HIGH*IHR=’+’ Z*IHR=’-’ /OVERLAY;
TITLE3 ’TREATMENT 1 MINUS TREATMENT 2’;
```

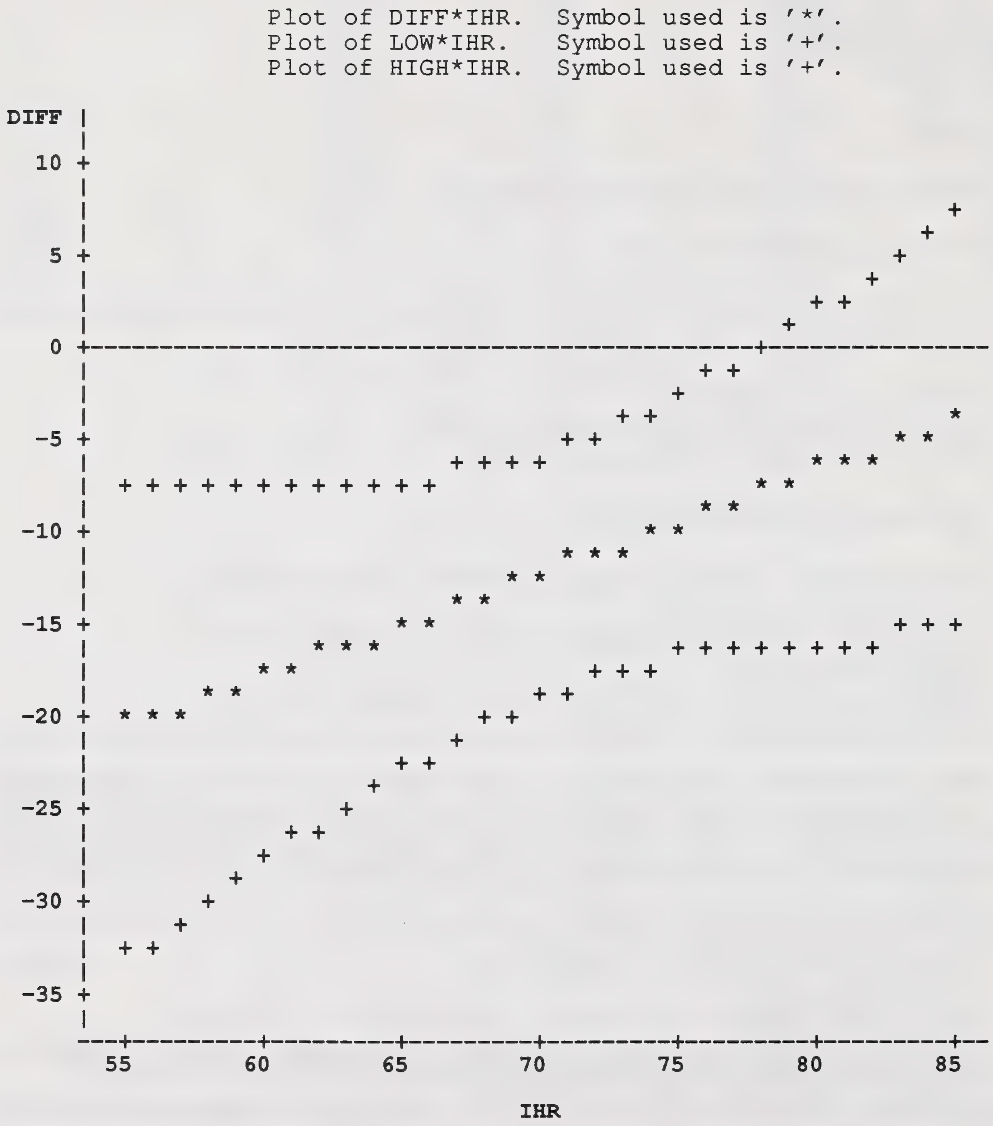


Figure 2.6 Confidence band about the difference of the models for EPRO one and two.

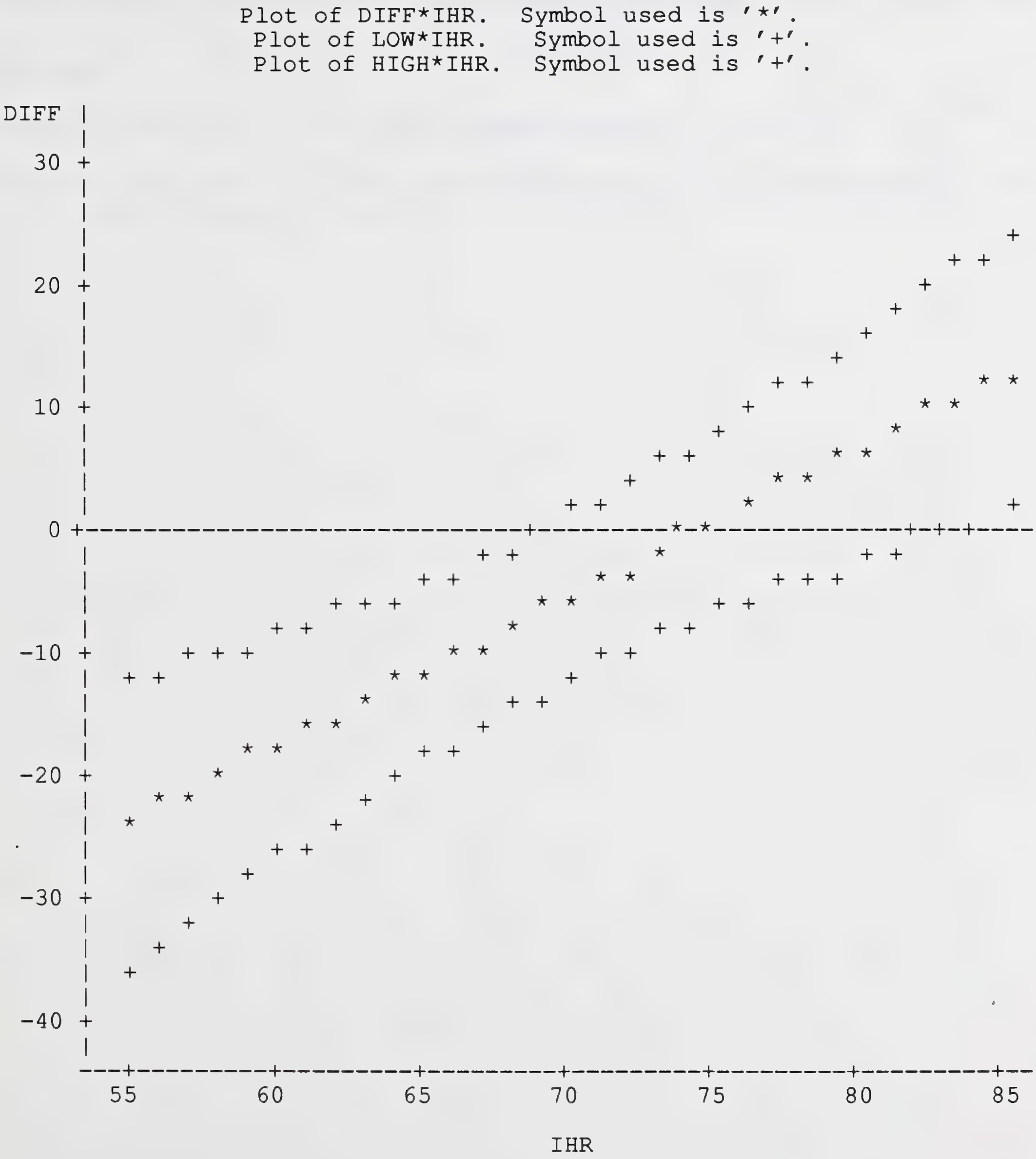


Figure 2.7 Confidence band about the difference of the models for EPROs one and three.

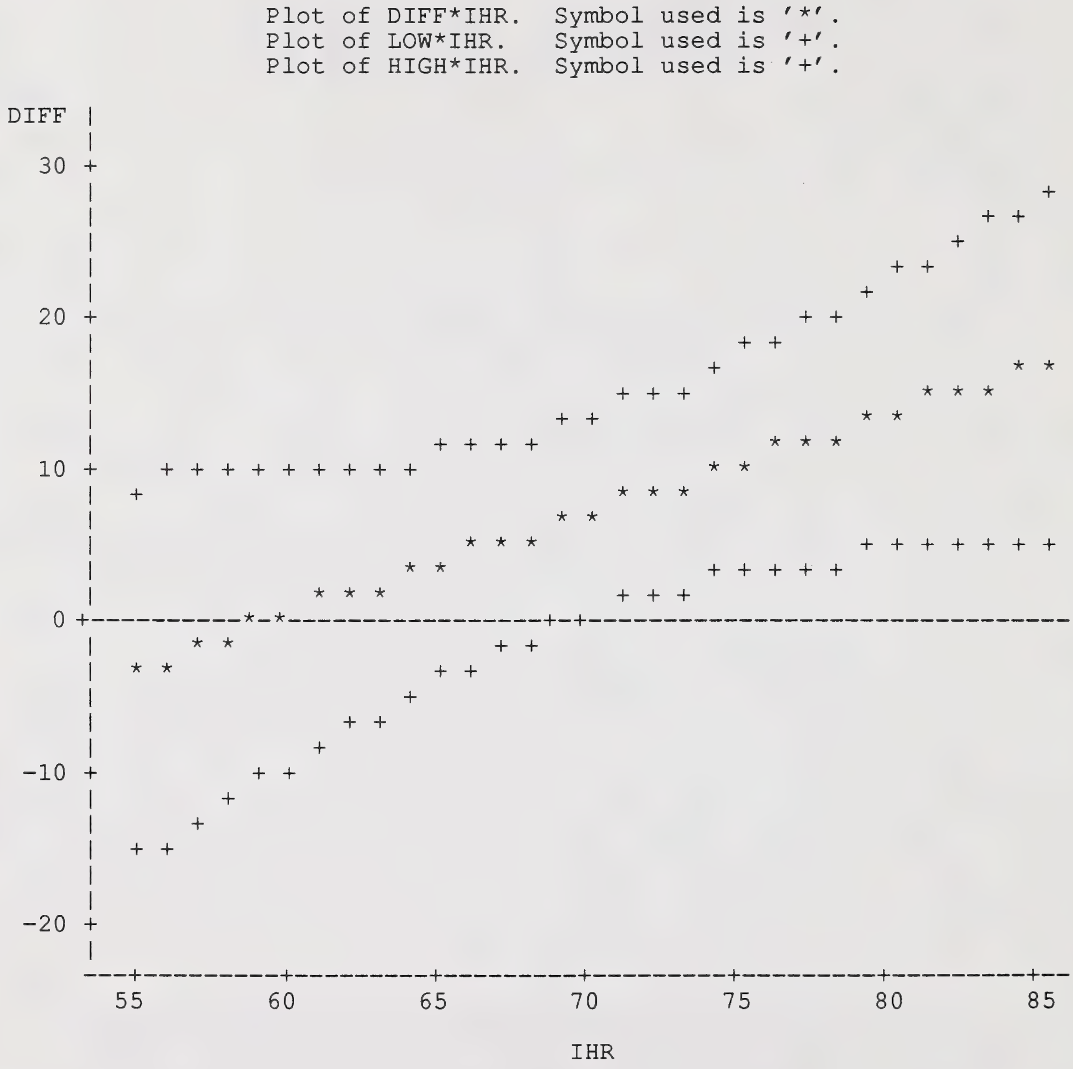


Figure 2.8 Confidence band about the difference of the models for EPROs two and three.

Summary

One must not blindly use computer code to do analysis of covariance. Several issues which must be addressed are (1) make sure the model adequately describes the data (linear, etc.), (2) check the equality of variances, (3) check the equality of slopes, (4) use the right model to make comparisons between the treatments, and (5) make sure you know what information is being obtained from your computer code.

References

Draper, N.R. and Smith, H. (1981). Applied Regression Analysis 2nd ed. New York: Wiley.

Milliken, G.A. and Johnson, D.E. (1984). Analysis of Messy Data. Vol I. Designed Experiments. New York: Van Nostrand Reinhold.

Two-Phase Regression

R.G. Weingardt
Programmer Analyst, Department of Animal Science
University of Alberta
Edmonton, Alberta

Abstract

Solutions to two-phase regression problems are generally found by using either the conventional method, which is a series of linear regressions, or alternately by using a nonlinear regression approach. The conventional method even in the simplest cases is time consuming prompting the researcher to turn to a quick solution via nonlinear regression. The use of one-shot nonlinear regression to solve two-phase regression problems often leads to incorrect solutions that look reasonable. To ensure that nonlinear regression reaches the correct solution bounds must be placed on one of the predictors.

Introduction

Two-phase straight-line regression is a statistical tool that has wide applications in many fields of research, and particularly in the areas of plant and animal physiology. A recent study (Nickerson, Facey, Grossman 1989) indicates that the two-phase procedure is only seldom used to its full potential on suitable data.

In this paper I will review the Nickerson *et al.* findings. Further, I will summarize the conventional least-squares approach to two-phase regression. Finally, I will discuss a modified nonlinear approach and give an example using the SAS procedure NLIN.

How Well is Two-Phase Regression Being Used?

In order to obtain an estimate of how appropriately two-phase regression is being used over a wide range of data Nickerson, Facey and Grossman (1989) studied all the papers published in *Physiological Zoology* from 1983 to 1987. They found that of the 31 papers that explicitly dealt with data suited to two-phase regression only one paper provided a statistically valid description of its data set.

The remaining 30 papers were categorized as follows: eight had no precise definition; nine fitted a regression line through only one section of the data; five did a separate regression of each section of the data but did not constrain the lines to meet between the two sets; seven did a visual estimate of the join point, (the join point is defined as the location of the intersection of the two straight lines which constitute the two-phase regression); and one researcher did a series of regressions based on preassigned x-values for join points and chose the combination which produced the smallest sum of squares. While some of the above methods seem to produce fairly reasonable estimates of the join points none of these produces a truly least-squares best fit estimate.

Conventional Method of Fitting a Two-Phase Straight-line Regression

Specification of the Model. This section uses equations and examples from Hudson (1966) and Nickerson, Facey and Grossman (1989). Please see these authors for a fuller description of the topic. For a discussion of the power of two-phase regression see Hinkley (1971).

Given a set of data as in Figure 1 and ignoring the distinction between sample and population parameters, the equations which state the conventional two-phase regression are

$$y_i = \text{INT1} + \text{SLP1} \times x_i + e_i, \qquad \text{if } x_i \leq \text{XJOIN, and}$$

(1)

$$y_i = \text{INT2} + \text{SLP2} \times x_i + e_i, \qquad \text{if } x_i > \text{XJOIN.}$$

In the best fit solution $\sum e_i^2$ is a minimum. Since the two regression lines meet at the join point whose abscissa is XJOIN, it follows that

(2) $$\text{INT1} + \text{SLP1} \times \text{XJOIN} = \text{INT2} + \text{SLP2} \times \text{XJOIN},$$

and the parameter of XJOIN can be calculated directly from equation (2),

(3) $$\text{XJOIN} = (\text{INT1} - \text{INT2}) / (\text{SLP2} - \text{SLP1}).$$

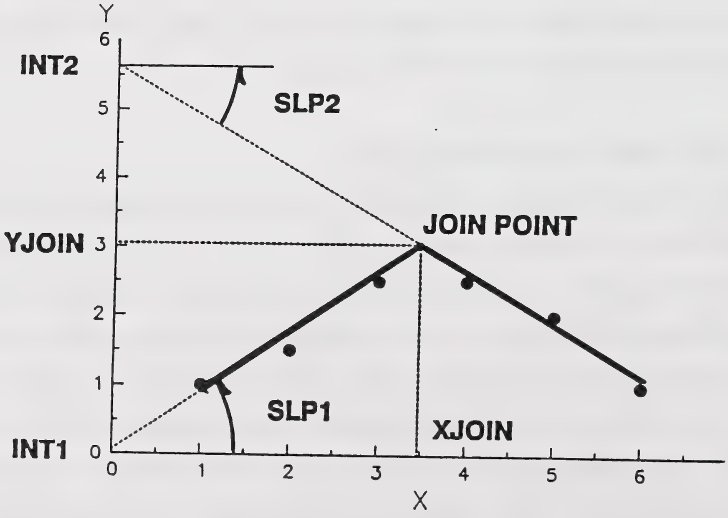


Fig. 1

Procedure for Determining the Best Two-Phase Model. First we order the x's from lowest to highest so that $x_1 \leq x_2 \leq \ldots \leq x_n$. We next split the data into two complementary sets corresponding to the two phases in which one set consists of the data points whose abscissas are x_1 to x_j while the other set has abscissas x_{j+1} to x_n. Simple linear regression is performed on each of the two sets. Since the minimum number of observations required to define a regression line is two then $2 \leq j \leq n\text{-}2$, where n is the number of observations. The maximum number of values for j is n-3 and is reduced by one for each duplicate x-value.

The coefficients from each of the n-3 (fewer if there are x-value duplicates) pairs of regressions are used to compute the corresponding XJOIN as in equation (3). How we proceed from here depends on whether or not $x_j < \text{XJOIN}_j < x_{j+1}$. If XJOIN for any j lies between the neighboring abscissas of its disjoint data sets then we have an ordinary least-squares (OLS) solution. The coefficients for such a j are as stated in equation (1) and the SSE of the combined equation is the sum of the SSEs of each of the two phases. If, *mirabile dictu*, all n-3 pairs of regressions produce OLS solutions then the best solution belongs to the pair with the smallest SSE.

For any value of j whose XJOIN does not lie within the exclusive range of x_j to x_{j+1} the solution is not valid. Here the regression lines have been generated with one or more data points placed in the wrong phase. This invalid pair of regression lines can be redefined by forcing the join point to equal x_j. This is called a constrained least-squares (CLS) solution. The equations for transforming an invalid OLS solution into a CLS solution are taken from Nickerson, *et al.* In the following definitions n_j and SS^j refer respectively to the number of observations and sum of squares in the first phase while $n_{j'}$ and $SS^{j'}$ refer similarly to the second phase.

(4) $\text{CLSINT1}=\text{INT1}-(s/t)\times(n_j\times SS^j{}_{xx})^{-1}\times\Sigma^j{}_{i=1}[(x_i-x_j)\times x_i]$,

(5) $\text{CLSSLP1}=\text{SLP1}+(s/t)\times(n_j\times SS^j{}_{xx})^{-1}\times\Sigma^j{}_{i=1}(x_i-x_j)$,

(6) $\text{CLSINT2}=\text{INT2}-(s/t)\times(n_{j'}\times SS^{j'}{}_{xx})^{-1}\times\Sigma^n{}_{i=j+1}[(x_i-x_j)\times x_i]$, and

(7) $\text{CLSSLP2}=\text{SLP2}+(s/t)\times(n_{j'}\times SS^{j'}{}_{xx})^{-1}\times\Sigma^n{}_{i=j+1}(x_i-x_j)$,

where

$$s = (\text{INT1} + \text{SLP1} \times x_j) - (\text{INT2} + \text{SLP2} \times x_j),$$

and

$$t=(n_j\times SS^j{}_{xx})^{-1}\times\Sigma^j{}_{i=1}(x_i-x_j)^2+(n_{j'}\times SS^{j'}{}_{xx})^{-1}\times\Sigma^n{}_{i=j+1}(x_i-x_j)^2.$$

For the CLS solution the sum of squares for error = $\Sigma^n{}_{i=1}\ (y_i-\hat{y})^2$. For any j-value the SSE of the constrained solution is greater than the SSE of the corresponding non-constrained solution be it valid (OLS) or invalid. It is, therefore, not necessary to perform CLS adjustments for any invalid solutions whose SSEs are greater than the smallest SSE among the valid solutions.

Even the simplest OLS solution requires up to (n-3)×2 simple linear regressions to obtain preliminary results and thereafter it requires either a great deal of hand calculation or some non-trivial programming to accumulate and process the various regression parameters. In the more typical problem involving some CLS solutions the time required to do the calculations could run into many hours while the risk of computational errors grows with the complexity of the problem. In those experiments for which regressions are required for each of many experimental units the calculation time can be very lengthy.

A Nonlinear Approach

Nonlinear procedures such as those contained in the SAS, SPSS-x, and BMDP packages are well suited to finding quick and accurate solutions to problems for which the SSE is a smooth function of each of the predictors. In two-phase regression the SSE is not a smooth function of XJOIN. A modified nonlinear procedure is therefore required. In this section I will employ the SAS procedure NLIN to illustrate this methodology. This procedure requires two adjustments not found in the conventional method: first, all the data are submitted to the program as one set rather than being run twice as two disjoint sets, and second, since the SSE as a function of XJOIN is not smooth at values XJOIN=x_i (Hudson 1966, p.1105), it is necessary to constrain the regression so that XJOIN does not equal any x_i.

In SAS's NLIN procedure it is possible to process both sets of data as defined in equation (1) simultaneously within one regression by using the IF statement to branch to either data set and its appropriate MODEL and DERivative statements (see SAS 1990, pp. 1162-3). For the sake of generality, I will redefine equation (1) so that only one model statement need be given. The single model statement can be used on a package lacking branching capability provided it has a derivative-free nonlinear regression routine.

The new parameters are defined as follows: YJOIN is the ordinate of the join point; XJOIN is the abscissa of the join point; AVESLOPE is the average of SLP1 and SLP2 from equation (1); and DELSLOPE is SLP2 minus AVESLOPE. The model may then be restated as

(8) $y_i = \text{YJOIN} + \text{AVESLOPE} \times (x_i - \text{XJOIN}) + \text{DELSLOPE} \times \text{ABS}(x_i - \text{XJOIN}).$

In the above equation ABS is a function that returns the absolute value of the expression in parentheses.

Example

The SAS program listed below does regression separately for two experimental units, Cow A and Cow B. The data, which are borrowed from Hudson (1966), permit accurate checking of results.

	x	1	2	3	4	5	6
Cow A y	1	2	4	4	3	1	
Cow B y	1	2	4	7	3	1	

For each data set the program searches for a minimum SSE with XJOIN constrained to lie successively within each of the three data intervals. The intervals are defined by the BOUNDS statement. One must be careful here to shrink the range between the x-value limits because SAS treats the hard inequality signs, < and >, as if they were soft inequality signs, ≤ and ≥. The MODEL statement is defined by equation (8).

Pure mathematicians will be shocked to see that DELSLOPE has a partial derivative, ABS(HOUR-XJOIN). The program is not affected by the lack of smoothness in the derivative because the point at which the slope changes abruptly, HOUR=XJOIN, has been precluded from any computations by the bounds put on XJOIN.

Table 1 contains the parameters (regrouped by cow ID) that were written by the PUT statement at the end of each of the three intervals. Comparing the results of Cow A with the curve in Figure 2a, one notes that the NLIN procedure identifies the minimum in the second interval and the local minimum in the third. Both are OLS solutions and are accurate to about eight digits. For Cow B the second and third interval solutions place XJOIN at the predefined boundaries. These are CLS solutions. The size of the error in the computed value of each of the parameters is a function of the size of the buffer around the x-values at the boundaries.

Table 1

Interval	Cow	XJOIN	YJOIN	SLOPE 1	SLOPE 2	SS Error
1	A	2.999990	4.315794	1.78949	0.92105	1.39475111
2	A	3.611111	4.750000	1.50000	-1.50000	0.33333333
3	A	4.193548	4.612903	1.10000	-2.00000	0.70000000
1	B	2.999990	5.368425	2.42107	-1.15789	11.57898127
2	B	3.999990	6.315802	1.92107	-2.78946	1.39477651
3	B	4.000010	6.315801	1.92105	-2.78950	1.39473684
	Hudson	4.000000	6.315789	1.92105	-2.78947	1.39473684

Listing of SAS Nonlinear Program

```
DATA raw;
INFILE 'a:twophase.dat';
INPUT  cow $ hour hormone KLICK;
PROC SORT; BY cow;
```

```
/*    ****  Solve for first interval    ****      */
PROC NLIN DATA=raw; BY cow;
PARMS YJOIN= 5  AVESLOPE= 0.1  DELSLOPE=-2     XJOIN= 2.5  ;
BOUNDS  2.00001 < XJOIN < 2.99999;
MODEL  hormone = YJOIN + AVESLOPE*(hour-XJOIN) + DELSLOPE*ABS(hour-XJOIN);
DER.YJOIN=1;  DER.aveslope=hour-xjoin;  DER.delslope=abs(hour-xjoin);
if hour > xjoin   then DER.xjoin=-aveslope-delslope;
if hour < xjoin   then DER.xjoin=-aveslope+delslope;
OUTPUT OUT=parsave1 PARMS=YJOIN AVESLOPE DELSLOPE XJOIN  SSE = SSE;
DATA new1; SET parsave1; FILE 'a:params1';
SLOPE1= AVESLOPE-DELSLOPE;   SLOPE2= AVESLOPE+DELSLOPE;
IF KLICK EQ 1   THEN  PUT COW  $  (XJOIN YJOIN) (11.6) (SLOPE1 SLOPE2) (10.5) SSE 13.8;

/*    ****  Solve for second interval    ****      */
PROC NLIN DATA=raw; BY cow;
PARMS YJOIN= 5  AVESLOPE= 0.1  DELSLOPE=-2     XJOIN= 3.5  ;
BOUNDS  3.00001 < XJOIN < 3.99999;
MODEL  hormone = YJOIN + AVESLOPE*(hour-XJOIN) + DELSLOPE*ABS(hour-XJOIN);
DER.YJOIN=1;  DER.aveslope=hour-xjoin;  DER.delslope=abs(hour-xjoin);
if hour > xjoin   then DER.xjoin=-aveslope-delslope;
if hour < xjoin   then DER.xjoin=-aveslope+delslope;
OUTPUT OUT=parsave2 PARMS=YJOIN AVESLOPE DELSLOPE XJOIN  SSE = SSE;
DATA new2; SET parsave2; FILE 'a:params2';
SLOPE1= AVESLOPE-DELSLOPE;   SLOPE2= AVESLOPE+DELSLOPE;
IF KLICK EQ 1   THEN PUT COW  $  (XJOIN YJOIN) (11.6) (SLOPE1 SLOPE2) (10.5) SSE 13.8;

/*    ****  Solve for third interval    ****      */      Similar to above
```

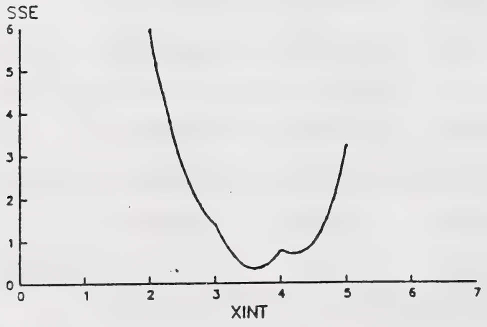


Fig. 2A

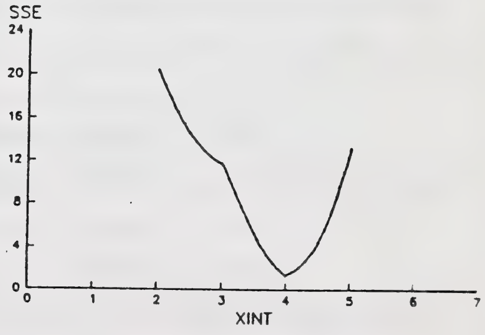


Fig 2B

Discussion

There are two scenarios regarding the general purpose of two-phase regressions. In the first a researcher may have many, say 20, data sets from which he wants to generate 20 join points and possibly

20 pairs of slope parameters to be used as data in subsequent analyses. At this point he may wish to test whether his data fit the two-phase model. The obvious hypothesis is, "Does the two-phase model effect a significant reduction in the error sum of squares over the simple linear model?" Unfortunately, the linear model is only one of many possible subsets of the two-phase model; the quadratic model is another. A test of significance of the R^2 change between the two-phase model and any of its subsets is a necessary but not a sufficient condition to warrant use of the two-phase model. One may use such a test (Overall and Klett, 1972, p. 422) as a guide, but in practice any set of data that has even a slight bend in the regression line will show a highly significant R^2 change in favor of the two-phase model over the linear model. Perhaps the best test is to closely "eye" the data's scattergram to see whether it looks two-phased rather than linear, quadratic or exponential.

In the second case a researcher has combined his data into one set and after deriving the regression parameters he may want to test whether either slope is different from zero. Nickerson *et al.* (1989) pp. 872-7 describe such a procedure. Further study may also be warranted for those regressions which produce a CLS solution. Rather than forcing XJOIN to equal some x_j, XJOIN may best be described as being the range x_j to x_{j+1} rather than a point. The solution to this model is given by Yeager and Ultsch (1989).

Conclusion

Unconstrained nonlinear regression will not in general provide the correct solution a two-phase regression problem. Nonlinear regression with bounds on the join point values provides accurate answers and does so much more quickly and easily than does the conventional approach.

References

Hinkley, D.V. 1971. Inference in two-phase regression. J. Am. Stat. Assoc. 66:736-742.

Hudson, D.J. 1966. Fitting segmented curves whose join points have to be estimated. J. Am. Stat. Assoc. 61:1097-1129.

Nickerson, D.M., D. E. Facey and G. D. Grossman, 1989. Estimating physiological thresholds with continuous two-phase regression. Physiol. Zool. 62(4):866-887.

Overall, J.E. and C. J. Klett, 1972. Applied multivariate analysis, New York NY. McGraw-Hill Book Company.

SAS Institute Inc. 1990. SAS/STAT User's Guide, Version 6, Volume 2. Cary, N.C.

Yeager, D.P. and G. R. Ultsch, 1989. Physiological regulation and confirmation: a BASIC program for the determination of critical points. Physiol. Zool. 62:888-907.

The Application of the MANOVA in Agriculture

L. A. Goonewardene
Alberta Agriculture, Beef Cattle & Sheep Branch
#204, 7000 - 113 Street
Edmonton, Alberta T6H 5T6

Abstract

The Multivariate Analysis of Variance Procedure (MANOVA) is useful when a large number of response variables are to be simultaneously analyzed and joint effects are to be determined. The test criteria that correspond to the F-test in a univariate analysis are: Wilks', Pillais, Hotelling-Lawley Trace and Roy's Greatest Root. The Wilks' is an exact F-test and most often used as a test criterion. The simultaneous response of the experimental units to all variables, considered as a single response, generally contains more information about the total effect of the treatment than does the series of responses considered individually. It is for this reason that a MANOVA should precede a univariate test under certain situations; an example is presented and discussed.

Introduction

The analysis of variance (ANOVA) is a widely used statistical procedure which partitions the variation into recognizable sources and tests for differences between the mean responses for the treatments. In the agricultural sciences, and animal science in particular, it is used in a Univariate Type of analysis where one or more main effects (X's) are tested in relation to a single response variable (Y). This is done usually by comparing the variance of the observed treatment means with an estimate of their expected variance, if there were no treatment differences. The response measure is assumed to be normally distributed with a constant variance.

The objective of this paper is to introduce the Multivariate ANOVA using the more widely used Univariate ANOVA as a basis and discuss briefly the application of the Multivariate ANOVA in agricultural experiments.

Multivariate ANOVA

Although the Univariate ANOVA is adequate for studying one dependent variable such as body weight, gain or grain yield at a time, there are situations when the dependents should be studied simultaneously. The SAS procedure has two methods to calculate ANOVA's, PROC ANOVA and PROC GLM. The main advantage in using PROC ANOVA is that it is much faster, but can only be used for balanced data. In the event that the data are unbalanced and least squares means are to be derived, PROC

GLM is the procedure to use. In addition, PROC GLM can be used for testing contrasts (weighted sums of the means) and will output a data set that contains residuals. Unfortunately GLM will not do a means separation such as Duncans, SNK or Tukeys on least squares estimates. The same procedures (ANOVA and GLM) can be used for Multivariate ANOVAS, referred to as a MANOVA.

The MANOVA is appropriate under many situations. Oftentimes there are a large number of response variables and it becomes difficult to look at each variable separately. Furthermore, when there is correlation between response variables one needs to identify the most important responses relative to the main effects. In instances where there is a high correlation between two responses, it could be argued that both responses are equally important as shown through a Univariate analysis whereas only one may be important. It is for this reason that a Multivariate ANOVA is considered necessary to precede a Univariate analysis (Littell et al. 1991), and only significant responses analyzed further by a Univariate approach. The simultaneous response of the experimental units to all variables, considered as a single response, generally contains more information about the total effect of the treatment than does the series of responses considered separately (Winer 1971).

Multivariate Tests

In a Univariate ANOVA, a main effect is usually tested by a F Statistic. This is derived by taking a ratio of the sums of squares, the one derived from the hypothesis being tested (Treatment Numerator) and the other derived from the unexplained or appropriate error term (Denominator). As Multivariate ANOVAs have more than one response variable (Y1 to Yn), the sums of squares are replaced by matrices of sums of squares and cross products.

The Multivariate linear model is written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}$$

where **Y** is a n x k matrix for k dependent/response variables, **X** is a n x m matrix of n observations on M dependents, **B** is a m x k matrix of regression coefficients and **e** is the n x k matrix of n random errors.

The SAS System has four functions which are used to determine significance synonymous with the F-test in a Univariate analysis. They are Hotelling-Lawley Trace, Wilks' Lambda criterion, Pillais Trace and Roy's Maximum Root. Simulation criteria have not been able to identify any one of the criteria as being superior (Littell et al 1991) although the Wilks Lambda criterion is used by many statisticians.

Example

Suppose we need to evaluate four swine feeds: a control (usually mixed on farm), new (mixed on farm) and Commercial A and Commercial B. The two dependent variables are: 1. How long does it take the pig to consume (TIME)?; and 2. How much is eaten (QUANT)? The experiment: Twenty pigs of the same breed, sex, age and stage of growth were randomly assigned to the four treatments C, N, A & B (C = Control, N = New, A = Comm A, B = Comm B).

SAS Code

```
DATA MANOVA;
* ONE WAY (R1) MANOVA - SWINE FEED TRIAL;
INPUT PIGNO 1-3 FEED $ 6-9 TIME 13-14
QUANT 19-22 .1;
CARDS;
;
PROC PRINT; TITLE 'PIG FEEDING TRIAL';
PROG GLM;
CLASS FEED;
MODEL TIME QUANT = FEED;
MEANS FEED;
MANOVA H = FEED;
OUTPUT OUT = RESIDS R = R1 R2;
RUN;
```

Output

The order in which these tables are presented is not necessarily the order in which the SAS output is obtained. This has been done to show you in what order the output should be read.

Table 1. Eigenvalues and Vectors

```
                         General Linear Models Procedure
                         Multivariate Analysis of Variance

     Characteristic Roots and Vectors of:  E Inverse * H, where
H = Type III SS&CP Matrix for FEED                        E = Error SS&CP Matrix

Characteristic                          Characteristic Vector    V'EV=1
    Root            Percent             TIME          QUANT

2.3315933541        96.27               0.01453729    0.50402729
0.0904626730         3.73               0.07588082    0.04497715
```

The characteristic roots are the eigenvalues. These represent linear functions of the dependent or response variables. The percent 96.27 represents the variation that each linear function explains. The vectors with large values show which of the two response variables are explained more by the model. This is sometimes very useful. In our example, QUANT = 0.504 is explained more than TIME = 0.014.

Table 2 shows the MANOVA analysis testing the null hypotheses of NO overall FEED effect.

Table 2 MANOVA analysis for FEED with test criteria

H = Type III SS&CP Matrix for FEED E = Error SS&CP Matrix

S=2 M=0 N=6.5

Statistic	Value	F	Num DF	Den DF	PR > F
Wilks' Lambda	0.275256	4.5302	6	30	0.0022
Pillai's Trace	0.782801	3.43	6	32	0.0100
Hotelling-Lawley Trace	2.422056	5.6515	6	28	0.0006
Roy's Greatest Root	2.331593	12.435	3	16	0.0002

Note: F Statistic for Roy's Greatest Root is an upper bound.
Note: F Statistic for Wilks' Lambda is exact.

There are four Multivariate tests. Using Wilks', which is an exact F-test with 6 degrees of freedom for the numerator and 30 for the denominator our probability of rejecting H_o is 0.0022. The conclusion is that FEED affects TIME and QUANT jointly. The same output may be obtained by using contrast statements as shown below:

```
FEED 1 0 0 -1,
FEED 1 0 -1 0,
FEED 1 -1 0 0;
```

The probability for the hypothesis of no overall FEED effect (H_o) can sometimes be smaller for joint effects compared to independent (Univariate) effects. Therefore, one could detect significant differences by viewing variables together that you would not detect by looking at variables individually. This is an advantage in the MANOVA approach.

Table 3 shows the mean and descriptive statistics generated due to inclusion of the PROC MEANS statement in the code.

Table 3 **Descriptive Statistics**

General Linear Models Procedure

Level of	TIME		
FEED	N	Mean	SD
ComA	5	26.2000000	4.14728827
ComB	5	27.2000000	3.42052628
Cont	5	25.2000000	3.11448230
New	5	27.8000000	2.58843582

Level of	QUANT		
FEED	N	Mean	SD
ComA	5	7.96000000	0.36469165
ComB	5	7.70000000	0.53851648
Cont	5	9.44000000	0.46151923
New	5	8.80000000	0.64807407

The data shows greater differences in QUANT compared with TIME. This will become evident with the Univariate tests.

Tables 4 and 5 show the Univariate tests for TIME and QUANT.

Table 4 **Univariate test for TIME with FEED as Main Effect**

Dependent Variable: TIME

Source	DF	Sum of Squares	F Value	PR > F
Model	3	19.60000000	0.58	0.6385
Error	16	181.20000000		
Corrected Total	19	200.80000000		

	R-Square	C.V.	TIME Mean
	0.097610	12.65137	26.60000000

Source	DF	Type I SS	F Value	PR > F
FEED	3	19.60000000	0.58	0.6385

Source	DF	Type III SS	F Value	PR > F
FEED	3	19.60000000	0.58	0.6385

Table 5 Univariate tests for QUANT with FEED as Main Effect

```
                              General Linear Models Procedure

Dependent Variable:  QUANT

Source                 DF    Sum of Squares        F Value       PR > F
Model                   3     9.51350000           12.01         0.0002
Error                  16     4.22400000
Corrected Total        19    13.73750000

                       R-Square      C.V.          TIME Mean
                       0.692520      6.062647      8.47500000

Source                 DF    Type I SS       F Value       PR > F
FEED                    3    9.51350000      12.01         0.0002

Source                 DF    Type III SS     F Value       PR > F
FEED                    3    9.51350000      12.01         0.0002
```

* Note: SAS gives you type I and III SS by default.

FEED is not a significant source of variation for TIME, P = 0.6385, but FEED is highly significant for QUANT. This same idea was noted by looking at the eigenvalues in Table 1. However, although QUANT is affected more by FEED, both TIME AND QUANT are affected jointly by FEED (Table 2).

There are option statements in the procedure that are helpful: the PRINTH and PRINTE, print the hypothesis and error matrices, respectively. The latter provides partial correlations from the sums of squares and cross products that are adjusted for all model effects. The SHORT option gives a condensed printout. If the data is balanced, PROC ANOVA may be used but PROC GLM should be used for unbalanced data. If CONTRAST statements are to be used, one should use PROC GLM. A factorial MANOVA will be discussed at the Workshop.

The MANOVA much like the ANOVA assumes that the response variables and residuals are normally distributed. The following code can be linked to the code provided on page 2, as a new OUTPUT data set has been created for variables R1 (TIME) and R2 (QUANT). The probabilities of rejecting a null hypothesis was 0.134 for R1 and 0.07 for R2 suggesting that both response variables did not deviate significantly from normality.

SAS Code

```
PROC UNIVARIATE Plot Normal data = RESIDS;
VAR R1 R2;
Title 'Check normality assumptions';
RUN;
```

Multivariate tests, although they lack power, are useful to detect joint effects. PROC ANOVA and PROC GLM can be used to perform MANOVAS. Although traditional approaches have looked at Univariate tests only, it may be better in certain in instances to look at MANOVA output and then decide on what Univariate variables are significant.

Acknowledgements

The author wishes to acknowledge the SAS Institute and authors who compiled the SAS notes on Multivariate Statistical Methods and Drs. R.C. Littell, R.J. Freund and P.C. Spector, from whose book some of the ideas for this paper were taken.

References

Hamer, R. 1988. Multivariate Statistical Methods: Practical applications and course notes. SAS Institute Inc., Cary, N.C.

Hand, D.J. and Taylor, C.C. 1987. Multivariate Analysis of Variance and Repeated Measures. A practical approach for behavioural scientists. Chapman and Hall, N.Y.

Littell, R.C., Freund, R.J. and Spector, P.C. 1991. SAS System for linear models (Ed. 3). SAS Institute Inc., Cary, N.C.

Milliken, G.A. and Johnson, D.E. 1984. Analysis of messy data, Volume 1 designed experiments. Lifetime Learning Publications, Belmont, California.

Winer, B.J. 1971. Statistical principles in experimental design. McGraw Hill Book Company, N.Y.

The Determination of Risk of Disease in Cattle Feedlots: A Case-Control Study

Casey Schipper
Animal Health Division
Alberta Agriculture
Edmonton, Alberta

Introduction

Shipping fever is a major cause of sickness in feedlot calves. The highest incidence of this disease occurs within six weeks after arrival at the lot. A variety of infectious agents as well as a multitude of management factors are now believed to be responsible for causing this disease. Factors such as mixing cattle from different sources, early silage feeding, processing, vaccination and water medication were associated with increased deaths and treatment costs in Ontario feedlots (1).

Objectives

The objective of this workshop presentation is to explain how simple cross-tabulation and logistic regression methods may be used to estimate the relative importance of contributing management factors (exposure factors) on the risk (probability of occurrence) of a given outcome variable (shipping fever).

Materials and Methods

Each line of this fictitious data set represents a partial health history of each of 2660 autumn weaned steer calves. The case-control study consisted of 284 calves affected with shipping fever and 2376 unmatched controls. The calves' individual eartag numbers were drawn by simple random selection from the records of a cattle feedlot in Alberta, housing approx. 6000 calves. The calves (units of analysis) in this sample population weighed between 180 and 340 kg (400 and 750 lbs), and entered the feedlot between Nov.15 and Dec 20, 1990.

For the purpose of this analysis the following exposure variables were considered:

1. SRC - About 10% of the sampled calves had been purchased through auction markets in Manitoba.
2. DH - About 36% of all calves required dehorning upon arrival.
3. DC - About 22% of the sample population underwent castration as well as dehorning at that time.
4. VAC - Vaccination against shipping fever was given to 36% of the calves.
5. SUR - Processing included either castration or dehorning or both.
6. BWT - Midpoint body weights selected were 438, 513, 588, 663 and 738 lbs.

The outcome variable "Shipping fever" was defined as a condition diagnosed by the feedlot penchecker/treatment person. Criteria for diagnosis were:

- body temperature greater than 40.5°C or (105°F)
- off feed, depression, abnormal breathing

Statistical Analysis

The **strength of association** between an exposure factor and disease is said to be the relative risk (RR), calculated as the ratio of the rate of disease in the exposed group to that in the unexposed group. The RR = 1 if there is no association between factor and disease. The further the RR deviates from 1 the greater is the strength of association.

The **odds ratio (OR)** is used in the analysis of case-control studies, since it closely approximates the value of the RR estimate. The OR is derived from the ratio a*d/b*c, where,

a = number of cases of disease in the exposed group

b = number of exposed controls

c = number of unexposed cases of disease

d = number of unexposed controls

The calculation of approximate 95% confidence limits on the OR is based on the natural log (ln) transformation of the approximate limits of ln(OR). One needs to calculate the variance of the OR, as follows (2)

$$\mathrm{var}\ (\ln OR) = (1/a + 1/b + 1/c + 1/d)$$

Since ln(OR) has a normal distribution in large samples the approximate 95% confidence limits for ln(OR) are,

$$\ln(OR) + 1.96 * (1/a + 1/b + 1/c + 1/d)^{0.5}$$

The approx. 95% lower and upper limits of ln(OR) are the antilog of the lower and upper limits of ln(OR) ie.

$$OR_L = OR \exp[-1.96 * \{\mathrm{var}\ (\ln OR)\})^{0.5}] \text{ and}$$

$$OR_U = OR \exp[+1.96 * \{\mathrm{var}\ (\ln OR)\})^{0.5}]$$

One needs to be cautious in interpreting risk of disease associated with exposure factors. The apparent or observed association may actually be due to another variable. The term **confounding** refers to the effect of an extraneous variable that wholly or partially accounts for the apparent effect of the study exposure factor. A confounding variable by definition is associated with the main independent variable studied and the outcome (disease) variable, but is not a consequence of exposure (3). Confounding variables can also be **determinants** of disease. If confounders are ignored in the analysis the real association between the exposure factor of interest and disease may be distorted. The control of confounding variables may be obtained by analytic methods (4).

The statistical method used for summarizing associations in multiple tables, is known as the Mantel-Haenzel (MH) technique. Different 2-by-2 tables showing number of cases and controls for each level of an exposure factor are constructed for each value level of the confounding variable. A limitation of the MH technique is the sheer numbers of 2-by-2 tables that need to be constructed to account for each category of perhaps many confounding variables (5). For instance in this study $2^8 = 256$ tables would have been required to obtain a Summary Odds Ratio for the risk of shipping fever.

The data set was created in Lotus 1-2-3. The Lotus data file was imported and converted into a epidemiologic data analysis program "EPI5". The **ANALYSIS** component of this program is designed to do frequencies, cross tabulations, means, graphs, and linear regression. The program computes OR values and their 95% confidence intervals and enables one to calculate Mantel and Haenzel Summary Odds ratios. The **STATCALC** module calculates statistics from tables entered from the keyboard. It also performs calculations for single and stratified 2-by-2 tables, determines sample size for cross-sectional, cohort and case-control studies and provides single and stratified trend analysis (6).

When the potential effects of a large number of suspected concurrent exposure factors need to be evaluated, multivariate logistic regression is commonly used. Multivariate modelling of relative risk functions has a 20-year history in epidemiologic research. The logistic regression approach is generally used in case-control studies of chronic disease. It is also useful in the analysis of disease risk in cattle feedlots (7). The technique allows one to examine the effect of one exposure or management factor while the values of other variables in the regression equation are held constant mathematically.

MULTLR is a Pascal microcomputer program, which performs interactive, menu-driven multiple logistic regression analyzed using conditional and unconditional maximum likelihood estimation methods. The program calculates most standard statistics and allows factoring of categorical or continuous variables by two distinct methods of contrast. A built-in, descriptive statistics option allows the user to inspect the distribution of cases and controls across categories of any given variable (8).

The binary outcome variable D (disease) is conditionally related to a series of independent regressors (discrete or continuous exposure variables) $x_1, x_2, ..., x_p$ through the formula (9):

$$\text{Prob}(D = 1| x_1, x_2, ..., x_p) = \{1 + \exp[-(a + b_1*x_1 + b_2*x_2 + ... + b_p*x_p)]\}.$$

where D denotes the presence (D = 1) or absence (D = 0) of disease, and x represents a set of p variables $x = (x_1, x_2, x_3,...x_p)$. The x variables represent any potential risk factor, confounding factor or any interaction term of interest. The b coefficients are parameters that represent the effects of the x's on the risk (probability) of disease.

The coefficient for any variable included in the logistic regression model depends on the entire set of variables included in the model. Thus a variable that by itself may be significantly associated with disease, may have a coefficient that renders it insignificant in multivariate analysis.

One should ensure that in judging the suitability of a fitted model the estimated effects of the variables included in the analysis make biological sense. Logistics regression models are used to obtain point (mean) estimates and suitable (95%) confidence limits for the probability (risk) of disease and unconditional or conditional odds ratios.

A detailed discussion on the rationale, methodology and interpretation of logistic regression functions may be found in suitable textbooks on the subject (10).

Results

The exposure factor with the greatest unconditional impact on the risk of shipping fever was the application of a surgical procedure, ie. either dehorning and castration (OR = 4.63, 95% CL = 3.55, 6.04). The source or origin of the cattle also had a significant impact on disease unconditionally (OR = 1.96, 95% CL = 1.33, 2.89). The weight of calves, categorized in 5 groups, was significantly associated with the risk of disease. Animals weighing between 400 and 475 lbs had a disease risk of 19.2% while only 1.8% of those weighing more than 700 lbs needed treatment for shipping fever. Dehorning by itself was not associated with an increased risk of illness ($p = 0.71$). Similarly, vaccination against Bovine Respiratory disease had no sparing effect ($p = 0.58$) on the risk of shipping fever in this data set (Table 1).

With respect to castration and dehorning, the source of calves was a confounder ($p = 0.002$), but vaccination and body weight were not. With respect to source however, body weight was a significant confounding variable ($X^2 = 110.2$ with 4 degrees of freedom).

The full logistic regression model only included dehorning/castration procedure, body weight, and the interaction term source*dehorning/castration as significant disease risk determinants ($p < 0.05$). The effect of source became insignificant when the remaining factors in the equation were controlled statistically. There was no interaction between surgery and the weight of calves (Table 2).

Taken in combination, the major factors influencing the probability of disease were light weight calves, dehorning/castration and the source of calves (Table 3).

Conclusion

The best calves to purchase from a disease control point of view based on this sample data set, are dehorned Alberta yearling steers. More than one half (64%) of all light weight Manitoba horned bull calves, entered into an Alberta feedlot under conditions prevalent in this data set would likely require treatment for shipping fever.

Table 1. Calf Sample Distribution of Shipping Fever Risk in Fictitious Alberta Feedlot (November 15, 1990 to January 30, 1991)

Variable ID	Variable Value	Variable Interpretation	Number (%) Exposed	Number (%) Diseased
Source	0	Alberta	2347 (91.6)	244(10.0)
	1	Manitoba	223 (8.4)	40 (17.9)
Dehorning	0	no dehorning	1706 (64.1)	185 (10.8)
	1	dehorning	954	99 (10.4)
Dehorning & Castration	0	No	2084 (78.3)	140 (6.7)
	1	Yes	576 (21.7)	144 (25.0)
Vaccination	0	No	1693 (63.6)	185 (10.9)
	1	Yes	967 (36.4)	99 (10.2)
Body Weight	438	lbs	950 (35.7)	182 (19.2)
	513	lbs	767 (28.8)	68 (8.9)
	588	lbs	605 (22.7)	22 (3.6)
	663	lbs	224 (8.4)	10 (4.5)
	738	lbs	114 (4.3)	2 (1.8)

Table 2. Maximum Likelihood Estimates of Logistic Parameters Relating Three Risk Factors to the Development of Shipping Fever in the Feedlot (November 13, 1990 to January 30, 1991).

Variable	Parameter (b)	Estimate b coeff	Std. Error of coeff	Odds Ratio	P-Value
x_0 Intercept	b_0	-2.252	0.095	-	0.000
x_1 Source	b_1	0.226	0.365	1.25	0.536
x_2 Castrate & dehorn	b_2	1.346	0.149	3.84	0.000
x_3 Body Weight	b_3	-1.899	0.307	0.15	0.000
$x_1 * x_2$	c^1	1.256	0.473	3.51	0.008
$x_2 * x_3$	c_2	0.275	0.416	1.31	0.509

Maximized log Likelihood L (b) = -779.38

Estimated Covariance Matrix

	b_0	b_1	b_2	b_3	c_1	c_2
b_0	0.009					
b_1	-0.007	0.133				
b_2	-0.009	0.007	0.022			
b_3	-0.008	-0.010	0.008	0.094		
c_1	0.008	-0.133	-0.016	0.011	0.224	
c_2	0.008	0.011	-0.018	-0.094	-0.044	0.1773

Table 3. A Summary of Maximum Likelihood Predicted Effects of Significant Exposure Factors and the Risk of Shipping Fever in an Alberta Feedlot (November 15, 1990 to January 30, 1991).

Outcome Variable	Values of Exposure Variables[1]			Predicted Treatment Rates Point Estimate (%)	Predicted Odds Ratios Point Estimate
Shipping Fever	SRC[2]	DC[3]	WT[4]		
	0	0	1	1.6	0.15 (6.67)
	1	0	1	1.9	0.19 (5.26)
	0	1	1	7.4	0.75 (1.33)
	0	0	0	9.5	1.00 (1.00)
	1	0	0	11.6	1.25
	1	1	1	25.9	3.33
	0	1	0	28.8	3.84
	1	1	0	64.0	16.91

[1] Values adjusted for the SRC*DC and DC*WT interactions.

[2] 0 = Alberta cattle 1 = Manitoba cattle

[3] 0 = no surgical stress 1 = dehorned and castrated on arrival

[4] 0 = calves weighing less than 550 lbs 1 = calves weighing 550 lbs and more

References

1. Martin SW, Holt JD, Meek HA, 1982(b). The Effect of Risk Factors as Determined by Logistic Regression on Health of Beef Feedlot Cattle. Proc. 3rd Int. Symp. Vet. Epidemiol. Econ. Sept. 1982, Arlington, VA.

2. Schlesselman J.J. Case-Control Studies, Design, Conduct and Analysis.(Monographs in Epidemiology and Biostatistics), Oxford University Press, New York, Oxford, 1982, p. 176.

3. Schlesselman J.J. Case-Control Studies, Design, Conduct and Analysis.(Monographs in Epidemiology and Biostatistics), Oxford University Press, New York, Oxford, 1982, p. 52.

4. Martin SW, Meek AH, Willeberg P, Veterinary Epidemiology, Principles and Methods. Iowa State University Press, Ames, Iowa. 1987, p. 134-135

5. Mantel N, Haenzel W, 1959 Statistical Aspects of the Analysis of Data for Retrospective Studies of Disease. J. Natl. Cancer Inst. 22:719-748.

6. Dean AD, Dean JA, Burton AH, Dicker RC, Epi Info, Version 5: a word processing, database, and statistics program for epidemiology on microcomputers. USD, Incorporated, Stone Mountain, Georgia, 1990.

7. Martin SW, Meek AH, Willeberg P, Veterinary Epidemiology, Principles and Methods. Iowa State University Press, Ames, Iowa. 1987, p. 296

8. Nelson Campos-Filho, BS and Eduardo L. Franco, A Microcomputer Program for Multiple Logistic Regression by Unconditional and Conditional Maximum Likelihood Methods, Amer. J. Epidemiology, (1989) 129 (2):439-444.

9. Schlesselman J.J. Case-Control Studies, Design, Conduct and Analysis.(Monographs in Epidemiology and Biostatistics), Oxford University Press, New York, Oxford, 1982, p.228.

10. Kleinbaum, DG, Kupper LL, Morgenstern H, Epidemiologic Research, Principles and Quantitative Methods. Lifetime Learning Publications, London, Singapore, Sydney,Toronto, Mexico City, 1982, p. 419.

Environmental Restoration and Statistics: Issues and Needs

Richard O. Gilbert
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352

This work was supported by the Environmental Restoration & Waste Management Division U. S. Department of Energy, Field Office Nevada, Las Vegas, NV

Abstract

Statistical analysis plays a vital role in environmental restoration (ER) activities. This paper is an attempt to show where statistics fits into the ER process. The statistician, as member of the ER planning team, works collaboratively with the team to develop the site characterization sampling design, so that data of the quality and quantity required by the specified data quality objectives (DQOs) are obtained. At the same time, the statistician works with the rest of the planning team to design and implement, when appropriate, the observational approach to streamline the ER process and reduce costs. The statistician will also provide the expertise needed to select or develop appropriate tools for statistical analysis that are suited for problems that are common to waste-site data. These data problems include highly heterogeneous waste forms, large variability in concentrations over space, correlated data, data that do not have a normal (Gaussian) distribution, and measurements below detection limits. Other problems include environmental transport and risk models that yield highly uncertain predictions, and the need to effectively communicate to the public highly technical information, such as sampling plans, site characterization data, statistical analysis results, and risk estimates. Even though some statistical analysis methods are available "off the shelf" for use in ER, these problems require the development of additional statistical tools, as discussed in this paper.

Introduction

The U.S. Department of Energy (DOE) has estimated that it will cost many billions of dollars to clean up its large volume of hazardous, radioactive, and mixed waste (DOE, 1990). As a consequence, the DOE is taking a close look at ways to reduce environmental restoration (ER) costs while still ensuring the scientific validity of the ER process. Achieving scientific validity begins by developing thorough plans for characterizing the waste site. These plans must specify both the quality and the quantity of site-characterization data required . That is, it is necessary to specify how much uncertainty in the nature and extent of contamination at the site can be allowed. Then efficient field sampling plans that will generate sufficient data of required quality can (and must) be developed.

In this paper, we discuss statistical issues and needs related to the planning and design of site characterization sampling studies and to the analysis of generated data. We begin by discussing how to deal with uncertainty (and cost) in the site-characterization process.

Dealing with Uncertainty in Site Assessments

It is impossible to remove all uncertainty about the nature, magnitude, and spatial patterns of contamination at a hazardous waste site, no matter how thorough sampling and measurement for the site characterization effort may be. We must therefore deal with the issue of uncertainty, which in turn means that we must deal with statistical concepts and methods that are related to the design of sampling programs so that we can quantify and/or reduce uncertainty. To deal with uncertainty requires careful planning even before any samples are collected. An important aspect of this planning process is the specification of data quality objectives (DQOs).

These DQOs are qualitative and quantitative statements that specify the quality of data that must be obtained. They are a tool to answer the questions: "What type and quality of data are needed to answer key questions?" and "How do we know when we have enough data?" (Neptune, et al. 1990). Specific DQOs are determined by the end use of the data and so are established to ensure that the data are both sufficient and of adequate quality for their intended use. Once DQOs are established by the ER planning team, they can be used by the statistician in developing a sampling design that will yield the necessary data at minimum cost.

The steps in establishing DQOs are as follows (Neptune, et al. 1990; EPA 1987a; EPA 1987b):

1. Carefully state the problem to be addressed or the decision to be made.
2. Identify the information required to select an appropriate course of action.
3. Articulate the specific role that data will play in selecting the course of action.
4. Specify the type of data needed.
5. Specify the way the data will be used.
6. Specify (by means of an iterative process that involves both the decision-maker and technical support staff) the degree of certainty desired in the conclusions to be derived from the data.
7. Optimize the sampling design for data collection to achieve the required degree of certainty in the conclusions at minimum cost.

Neptune, et al. (1990) have illustrated the procedure with a case study. The question addressed in the case study is whether a site at which railroad ties and creosote-soaked timbers were stored and burned posed an unacceptable risk to site workers and visitors as a result of exposure to polyaromatic hydrocarbons (PAH) in soil. After carefully considering the human health consequences and consulting with the toxicologist and site engineers, the project manager assigned acceptable decision error rates for various risk levels and associated levels of average PAH soil concentrations. Then a soil sampling plan was developed that was expected to achieve these DQOs at minimum cost. The sampling plan specified the number of composite soil samples that should be collected to estimate average soil PAH concentrations with sufficient precision to achieve the DQOs, i.e., such that the specified decision error rates were not exceeded.

Dealing with Costs Using the Observational Approach

One lesson learned from ten years of experience of the U.S. Environmental Protection Agency (EPA) with conducting clean-up actions under the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) is that the clean-up process must be streamlined to avoid high costs and long delays in conducting remedial actions (EPA 1989a). The DOE and the EPA are currently evaluating the "Observational Approach" as a framework for streamlining the clean-up process while managing the uncertainty inherent in site assessment. Overviews of the method have been published by Smyth and Quinn (1991) and by Myers and Gianti (1989). The advantages and limitations of the method are discussed by Peck (1969). Brown, et al. (1989) discuss the application of the method to the remediation of hazardous waste sites. The method is a way of initiating remedial action at a waste site without full characterization of the nature and extent of contamination. The observational approach is intended to reduce costs by accepting greater uncertainty and allowing earlier selection of a remedial action approach based on probable conditions at the waste site.

The uncertainty in knowledge about conditions at the site is taken into account by making contingency plans for handling deviations from probable conditions if they occur during remedial action. In essence, the observational approach requires conducting thorough up-front planning to identify uncertainties and determine both possible and probable conditions at the site. The remedial action program is designed for probable conditions, but contingency plans are prepared in case deviations from the probable conditions occur during remedial action. As indicated by Smyth and Quinn (1991), the DOE has endorsed the concept (DOE, 1990), and the EPA has endorsed an equivalent approach (EPA, 1989b).

Statistical Needs in Environmental Restoration

It is generally recognized that statistical methods should be used in ER projects. However, a number of problems associated with using statistical methods must be addressed. The following is a short discussion of some of these problems.

Understanding the Role of the Statistician

A statistician can contribute to the ER process in many ways, including helping to define the DQOs, developing sampling plans, conducting data analyses, reviewing analytical laboratory protocols and quality control procedures, developing strategies for risk assessment, quantifying the uncertainty of model predictions by using uncertainty and sensitivity analyses, developing graphics for communicating data and results, reviewing draft reports, and contributing to peer-reviewed papers and reports. These contributions will be most effective when the statistician is a member of the ER planning team, preferably from the very beginning of the ER planning process.

Developing Data Quality Objectives

Data quality objectives must be specified as part of the planning process. If this is not done, no one will know when to stop collecting data, and a lot of unnecessary or inappropriate data are likely to be the result. This idea of specifying a priori the quantity of data that is needed is a familiar concept to statisticians. For example, the method statisticians use to determine the number of samples required for estimating a mean involves specifying the required accuracy of the estimated mean and the confidence required in achieving that accuracy (Gilbert, 1987, pp. 30-42). The entire ER planning team needs to understand and support the use of DQOs and to take part in determining what those DQOs should be. The DQO approach will provide the information needed to develop efficient sampling plans and associated statistical analysis methods.

Reducing Cost

Ways must be found to reduce the cost of ER. Efficient planning via the Observational Approach and the specification of DQOs are two important tools for that purpose. But other methods are also available. For instance, the compositing of several field samples into one thoroughly mixed sample, which is then subsampled for analysis, can reduce analytical costs when the method is applicable (Gilbert, 1987; Bolgiano, et al. 1990). For example, concentrations of composite soil samples were used to evaluate the need for further removal of soil at sites contaminated with dioxin in the State of Missouri (Exner, et al. 1985). The use of in situ detectors in place of a portion of the environmental samples can also sometimes reduce costs. For example, in situ spectrometry was used in the United States on the Nevada Test Site, a nuclear weapons testing area, to estimate the spatial distribution and total inventory of the important anthropogenic radionuclides in the surface soil (McArthur, 1991). When in situ detectors are used to estimate environmental concentrations, it is usually necessary to quantitatively relate the in situ detector readings to concentrations in samples collected at the same locations. This quantification requires a statistical analysis of both the in situ and the sample data collected using a valid statistical design according to established DQOs.

Coping with Spatial Variability

Contaminant concentrations can vary greatly over space. Site characterization usually implies estimating what contaminants are present, where they are located, and their concentrations. If contaminant concentrations are not too heterogeneous over space, geostatistical techniques, such as kriging, can be used to estimate the spatial pattern by taking into account spatial correlations (Flatman, 1984; Gilbert and Simpson, 1985). The use of geostatistics for evaluating the attainment of clean-up standards is discussed in EPA (1989c). However, the presence of hot spots is another complicating problem. Although simple methods are available for determining the spacing between points on a sampling grid required to detect hot spots with specified probability (Gilbert, 1987), the number of sampling locations required can be prohibitively large. Sometimes it is possible to reduce variability by compositing and mixing samples. Also, in situ detector measurements can be less variable than those of single small samples because the in situ measurements measure relatively large volumes of soil. However, such smoothing out of sample spatial variabilities can also hide hot spots.

Characterizing Highly Heterogeneous Waste

Many hazardous waste sites contain heterogeneous materials; e.g., concrete, clothing, liquids, bottles, tires, paper, and shredded autos may all occur at a site. These materials may be packed in barrels or lying loose. A given barrel may contain many different types of waste, unknown unless the barrel is opened and inspected, which is an expensive and possibly dangerous operation. The EPA is searching for techniques for obtaining representative samples of debris from hazardous-waste sites. As stated and discussed by Rupp (1990), the problems include 1) obtaining a representative sample from a mix of materials of various sizes and compositions, 2) characterizing the contamination of large items in a way that has meaning for a health risk assessment, and 3) subsampling from mixtures of large objects to produce the small-volume samples required by the analytical laboratory. The basic question is whether a defined unit of material, e.g., a barrel, contains areas of contamination that exceed action levels. A related problem is how to reduce the number of barrels that need to be opened and characterized. A sampling approach, such as acceptance sampling (Schilling, 1982), might be feasible to resolve the latter problem. But generally, the solution of these problems should be a team effort, wherein DQOs are established first, followed by studies to determine sampling and inspection approaches that meet the DQOs.

Coping with "Less-Than" Data

Frequently the concentration of a contaminant in a field sample cannot be quantified, in which case it may be reported as a nondetectable or "less-than" value. When such less-than values are present, it is difficult to obtain valid estimates of important parameters, such as mean concentrations, or to conduct valid statistical tests to identify changes in concentrations over time or determine compliance with clean-up standards. The common practice of replacing less-than values with zeros or other fabricated values can lead to highly misleading results. To avoid this problem, it is necessary to use special statistical methods (Helsel, 1990; Gilbert, 1987). The need for additional statistical methods for such cases is discussed by Lambert, et al. (1991) who also introduce new tools (the "probability of acceptance" and the "probability of detection") to describe which measurements and field concentrations are detectable.

Using Uncertainty and Sensitivity Analyses

The assessment of risks associated with various clean-up scenarios and technologies requires the use of environmental transport and risk models, the predictions of which are often highly uncertain. Yet this uncertainty may not be explicitly taken into account when formulating DQOs and making decisions. User-friendly computer codes are available for conducting Monte Carlo uncertainty analyses (Iman and Shortencarier, 1984) and sensitivity analyses (Iman, et al. 1985) to quantify the uncertainty in model predictions and to identify model parameters that have a big impact on model predictions. Use of these methods may be considered if DQOs require precise estimates of uncertainty and when decisions must be made about which model parameters should be refined to reduce uncertainty. The use of uncertainty and sensitivity analyses has the additional benefit of clarifying the sources of uncertainty in the model and model parameters. This identification process also helps develop a more thorough understanding of the uncertainties present in the system and where they reside. IAEA (1989) has described procedures for evaluating the reliability of predictions made by environmental transport models, including uncertainty and sensitivity analyses and validation studies. Finkel (1990) has discussed uncertainty in risk management for decision-makers.

Using Nonparametric Statistical Tests

The standard assumption that underlies many statistical tests of hypotheses is that the data are normally (Gaussian) distributed, which is, however, not usually the case with waste-site data. In such situations, nonparametric statistical tests should be considered. For example, consider the problem of testing for attainment of risk-based or background-based soil clean-up standards at a remediated waste site. The nonparametric Wilcoxon Rank Sum test and Quantile test (Gilbert and Simpson, 1990) can be used for the background-based case. For the risk-based case, nonparametric tests for proportions based on the binomial distribution can be used (EPA, 1989c). Nonparametric tests are not a cure-all. For example, the data must be uncorrelated in order for the nonparametric test results to be valid, the same requirement as for standard parametric tests. However, nonparametric tests can be more powerful (i.e., have a smaller false-negative error rate) than parametric tests when the assumption of normality is not valid. Also, some nonparametric tests can be used even when a moderate number of less-than values are present (Gilbert, 1987), if all less-than values are less than the smallest detected value. The Quantile test can be used even when a large proportion of the data are less-than values. Regulators and the planning teams for ER need to become familiar with nonparametric tests and their advantages and disadvantages.

Communicating with the Public

The public supports ER with tax dollars. Every effort should therefore be made to effectively communicate the plans, methods, results, and implications of results to the public in forms that the average person can understand. This is a formidable challenge. Statistical graphics and geographical information systems (GIS) are tools that can contribute to this communication effort (Tzemos, et al. 1991; Dangermond and Harnden, 1990). For example, a GIS can be used to integrate hazardous-waste site data with associated geographic information via map overlays. There is also potential for integrating statistical analyses and modeling of spatial data into GIS software, although present capabilities are limited (Bailey, 1990). This is an area where more work is needed to assess what is required and what the costs might be to develop the methodology. Also, statisticians can take mysticism out of statistical techniques by avoiding jargon, going back to first principles, and using intuitive descriptions and examples.

Conclusions

The problems of cost and uncertainty associated with ER can be tackled using data quality objectives, the observational approach, and statistical designs and analyses developed by the statistician in collaboration with other members of the ER planning team. A number of statistical tools are currently available, but the development of additional tools is needed. This paper has described a few of the issues and needs related to the application of statistics to site characterization and ER at hazardous waste sites.

Acknowledgments

This work was supported by the Environmental Restoration & Waste Management Division, U.S. Department of Energy, Field Office Nevada, Las Vegas, Nevada. The Pacific Northwest Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RLO 1830.

References

Bailey, T. C. 1990. "GIS and Simple System for Visual, Interactive, Spatial Samples," The Cartographic Journal 27:79-84.

Bolgiano, N. C., Patil, G. P. and Taillie, C. 1990. "Spatial Statistics, Composite Sampling, and Related Issues in Site Characterization with Two Examples," Proceedings of the Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs, pp. 79-117. U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, Statistical Policy Branch, Washington, D.C.

Brown, S. M., Lincoln, D. R. and Wallace, W. A. April 1989. Application of the Observational Method to Remediation of Hazardous Waste Sites, CH2MHILL, P.O. Box 91500, Bellevue, WA, 98009.

Dangermond, J. and Harnden, E. 1990. "Map Data Standardization - A Methodology for Integrating Thematic Cartographic Data Before Automation," ARC News 12:16-19.

DOE. 1990. Environmental Restoration and Waste Management Five-Year Plan Fiscal Years 1992-1996, DOE/S-0078P, U. S. Department of Energy, Washington, D.C.

EPA. March 1987a. Data Quality Objectives for Remedial Response Activities: Development Process, Office of Emergency and Remedial Response, U. S. Environmental Protection Agency, Washington, D.C.

EPA. March 1987b. Data Quality Objectives for Remedial Response Activities: Example Scenario, Office of Emergency and Remedial Response, U. S. Environmental Protection Agency, Washington D.C.

EPA. January 1989a. RI/FS Improvements Phase II, Streamlining Recommendations," OSWER Directive No. 9355.3-06, Office of Emergency and Remedial Response, U. S. Environmental Protection Agency, Washington, D.C.

EPA. 1989b. RI/FS Streamlining, OSWER Directive No. 9355.3-06, U. S. Environmental Protection Agency, Washington, D.C.

EPA. 1989c. Methods for Evaluating the Attainment of Cleanup Standards, Volume 1: Soils and Solid Media, EPA 230/02-89-042, U. S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, Statistical Policy Branch, Washington, D.C.

Exner, J. H., Keffer, W. D., Gilbert, R. O., and Kinnison, R. R. 1985. "A Sampling Strategy for Remedial Action at Hazardous Waste Sites: Clean-up of Soil Contaminated by Tetrachlorodibenzo-p-Dioxin," Hazardous Waste & Hazardous Materials 2:503-521.

Finkel, A. M. 1990. Confronting Uncertainty in Risk Management, A Guide for Decision-Makers, Center for Risk Management, Resources for the Future, **Washington, D.C.**

Flatman, G. T., 1984. "Using Geostatistics in Assessing Lead Contamination Near Smelters," Environmental Sampling for Hazardous Wastes, ACS Symposium Series 267. American Chemical Society, Washington, D.C., pp. 43-52.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold, New York, NY.

Gilbert, R. O. and Simpson, J. C. 1985. "Kriging for Estimating Spatial Pattern of Contaminants: Potential and Problems," Environmental Monitoring and Assessment 5:113-135.

Gilbert, R. O. and Simpson, J. C. 1990. "Statistical Sampling and Analysis Issues and Needs for Testing Attainment of Background-Based Cleanup Standards at Superfund Sites," Proceedings of the Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs, pp. 1-19, U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, Statistical Policy Branch, Washington, D.C.

Helsel, D. R. 1990. "Less Than Obvious: Statistical Treatment of Data Below the Detection Limit," Environmental Science and Technology 24:1766-1774.

IAEA, 1989. Evaluating the Reliability of Predictions Made Using Environmental Transfer Models, IAEA Safety Series No. 100, International Atomic Energy Agency, Vienna.

Iman, R. L. and Shortencarier, M. J. 1984. A FORTRAN 77 Program and User's Guide for the Generation of Latin Hypercube and Random Samples for Use With Computer Models, NUREG/CR-3624, Sandia National Laboratory, Albuquerque, NM.

Iman, R. L., Shortencarier, M. J. and Johnson, J. 1985. A FORTRAN 77 Program and User's Guide for the Calculation of Partial Correlation and Standardized Regression Coefficients, NUREG/CR-4122, Sandia National Laboratories, Albuquerque, NM.

Lambert, D., Peterson, B. and Terpenning, I. 1991. "Nondetect, Detection Limits, and the Probability of Detection," Journal of the American Statistical Association 86:266-277.

McArthur, R. D. 1991. Radionuclides in Surface Soil at the Nevada Test Site, DOE/NV/10845-02, Water Resources Center Publication # 45077, Desert Research Institute, Las Vegas, NV.

Myers, R. G. and Gianti, S. J. 1989. "The Observational Approach for Site Remediation at Federal Facilities," Superfund '89 Proceedings of the 10th National Conference, November 27-28, Washington, D.C.

Neptune, D. Brantly, E. P., Messner, M. J. and Michael, D. I. 1990. "Quantitative Decision Making in Superfund: A Data Quality Objectives Case Study," Hazardous Materials Control 3:18-27.

Peck, C. E. 1969. "Advantages and Limitations of the Observational Method in Applied Soil Mechanics," Geotechnique 19:171-187.

Rupp, G. 1990. Debris Sampling at NPL Sites, Draft Interim Report, Environmental Research Center, University of Nevada-Las Vegas, Las Vegas, NV

Schilling, E. G. 1982. Acceptance Sampling in Quality Control, Marcel Dekker, New York, NY.

Smyth, J.D. and Quinn, R.D. 1991. The Observational Approach in Environmental Restoration, PNL-SA-18817, Presented at the 1991 American Society of Civil Engineers National Conference on Environmental Engineering, July 8-10, 1991, Reno, NV. Pacific Northwest Laboratory, Richland, WA.

Tzemos, S., Evans, B. J., and White, M. E. 1991. Developing a GIS to Facilitate Data Analysis for Environmental Restoration of a Large Waste Site, PNL-SA-19158, Presented at the 11th Annual Environmental Systems Research Institution User Conference, May 20-24, 1991, Palm Springs, CA. Pacific Northwest Laboratory, Richland, WA.

Some Problems in Statistical Consulting

G. C. Kozub
Agriculture Canada Research Station
Lethbridge, Alberta, T1J 4B1

Abstract

The role of consulting statisticians in agricultural research has changed in recent years, with collaboration and advising on research studies now being their major activity. The trend for researchers to undertake some or all of their own statistical work has merit if they have a good knowledge of statistical methods and analysis techniques, but can lead to lower research standards otherwise. Some guidelines for the handling of statistical problems are discussed.

Introduction

Corresponding to advances that have taken place in computer technology in recent years, there has been an increase in the number of users of statistical methods in agricultural research. These include people with considerable formal training and experience in statistics, as well as those whose main expertise is in another field, but within which statistics plays an important part. Before the arrival of microcomputers, the bulk of statistical work in agricultural research was handled using large mainframe computer systems, and often through a professional statistician or individual who had sufficient statistical knowledge to be designated as the "local statistician".

Today, with the proliferation of computer terminals, microcomputers and statistical software, increased amounts of information to be analyzed, and limited numbers and availability of consulting statisticians, more people with varying degrees of statistical training are carrying out their own statistical work. This has led to a change in the role of the statistician, and the real danger of misapplication of statistical methods by researchers with limited statistical backgrounds. I would like to provide an overview of statistical consulting at an agricultural research station and indicate some general guidelines for and problems encountered when handling statistical problems.

Statistics at the Lethbridge Research Station

The Lethbridge Research Station is an agricultural research station in Agriculture Canada that has research programs in Crop Science, Livestock, and Soil Science sections. About 65 scientists carry out research studies, and statistical considerations play an important part in many of these. The Station has 3 statisticians and a data analyst who work with the scientists on statistical aspects of their studies. Most of the statistical processing is carried out using SAS (SAS Institute, Inc.) software on a VAX system, but other statistical software, on the VAX and microcomputers, is also used.

The statisticians have limited to major input in statistical aspects of specific research studies. The statistician's role can be as a collaborator, advisor, or to provide a service. As a collaborator, the statistician plays an essential part in the study, working closely with the scientist and carrying out most of the statistical work. As an advisor, the statistician is consulted about various aspects of the design and analysis, but the major burden of the work is done by the scientist. When providing a service, entire routine tasks are taken on by the statistician, and there is a low level of personal involvement in the study. In recent years, the service role has diminished to a low level, with researchers performing routine analyses and the consulting statistician spending most of his time as a collaborator or advisor. Genuine collaboration, where the statistical problem is challenging and the scientist and statistician pool their talents and expertise to produce high quality research, is the most rewarding role for the consulting statistician.

Although it is beneficial for researchers to use statistical methods, this can result in inadequate analyses, misapplication of statistical methods, and inefficient use of computer resources if they do not have a good knowledge of statistical methods and analysis techniques. Researchers need to use statistical methods frequently to retain their skills and develop the ability to recognize statistical problems and processes that they can handle themselves, and those for which professional assistance is required. In addition, the researcher must consider the time requirement for statistical analysis so that it does not take too much time away from his own discipline. Easy to read documentation on the use of statistical software combined with one-to-one assistance from statistical personnel is helpful for improving their effectiveness and self-interest. Even when advice is obtained from a professional statistician, inappropriate analyses can occur if the statistician does not have sufficient information on or time to digest a problem, often as a result of work pressures and a backlog of requests.

Some guidelines for handling statistical problems

When working with scientists on a research study, the statistician can enter the study at the planning, analysis, or interpretation, presentation, and publication stages. Although it is ideal to enter during the planning and experimental design stage, and be fully involved at all subsequent stages, the statistician often enters at the analysis or later stages. It is extremely important that the statistician discuss the problem in detail with the scientist to get an accurate and complete description and understanding of those parts of the research that have an implication on the design and statistical analysis. The objectives should be clearly formulated, and plans for the design and analysis outlined. Good documentation of key discussions with the scientist at all stages of the study are valuable for the efficient handling of problems. For complex projects, a written summary of the statistician's understanding of the project and the agreed on course of action should be sent to the scientist.

Prior to doing a detailed statistical analysis of the data, considerable time should be spent carefully scrutinizing the data, both visually and using computer summaries and plots, to check for entry errors, discrepant values, heterogenous variance, non-linearity, and so on. Initial analyses on the data and another review of the objectives with the scientist will clarify the analyses that are really important and the approach to take for obtaining these. Detailed consideration has to be given as to whether data transformations are necessary, the choice of statistical techniques to be used when there are different possibilities, and whether standard or specialized statistical methods are needed.

Commonly used statistical methods found in standard texts like Bliss (1967), Cochran and Cox (1957), Draper and Smith (1981), Fleiss (1981), Milliken and Johnson (1984), Snedecor and Cochran (1980), Steel and Torrie (1980) and Winer (1962) can be used in the majority of research problems and are more likely to be understood by researchers. However, there are times when nonstandard or highly complex statistical procedures are needed to meet the objectives. For example, at an agricultural research station plant breeding studies may be carried out to evaluate many varieties and lines for various characters such as yield and plant height. It is often of interest to determine groupings of the varieties and lines that respond similarly at different locations, over years, exhibit a similar growth response, and so on. A joint regression technique (Finlay and Wilkinson 1963), which was expanded to include clustering (Lin and Thompson 1975), is useful for investigating complicated two factor interaction structures such as these.

Once the statistical analyses have been obtained, the output should always be checked very carefully to ensure that the analyses are correct and that the statistics produced correspond to what the raw data indicate. Statistical software packages can give incorrect or inappropriate results that only become apparent with close examination of the output. The final analyses should be summarized using tables and diagrams that are fully informative of the data from which they have been derived, and carefully interpreted with the original objectives in mind. A written report of the statistical methods used and results obtained should be prepared for the scientist so that they are clearly understood and properly reported. Drafts and galley proofs of the publication must be reviewed carefully to ensure that what is published is what was intended.

Summary

The ways in which statistical design and analysis problems are handled has changed in recent years. With the increased availability and use of computers, consulting statisticians are assuming primarily collaborative or advisory roles, and more researchers are carrying out their own statistical work. Although this has merit, the use of statistical methods by individuals with a limited knowledge of statistical methods and analysis techniques can lead to insufficient or incorrect analyses, and important and interesting features of data may not be brought out. It is important for researchers to develop skills to identify problems where the appropriate statistical analysis is outside their knowledge or ability. Some guidelines for the handling statistical problems are discussed.

References

Bliss, C. I. 1967. Statistics in biology, Vols. 1,2. McGraw-Hill Book Co., Toronto.

Cochran, W. G. and G. M. Cox. 1957. Experimental designs, 2nd ed. John Wiley and Sons, Inc., Toronto.

Draper, N. R. and Smith, H. 1981. Applied regression analysis, 2nd ed. John Wiley and Sons, Inc., Toronto.

Finlay, K. W. and Wilkinson, G. N. 1963. The analysis of adaptation in a plant breeding programme. Aust. J. Agric. Res. 14: 742-754.

Fleiss, J. L. 1981. Statistical methods for proportions, 2nd ed. John Wiley and Sons, Inc. Toronto.

Lin, C. S. and Thompson, B. K. 1975. An empirical method of grouping genotypes based on a linear function of the genotype-environment interaction. Heredity 34: 255-263.

Milliken, G. A. and Johnson, D. E. 1984. Analysis of messy data, Vol. 1. Van Nostrand Reinhold Co., New York, N.Y.

Snedecor, G. W. and W. G. Cochran. 1980. Statistical methods, 7th ed. Iowa State Univ. Press, Ames, Iowa.

Steel, R. G. D. and Torrie, J. H. 1980. Principles and procedures of statistics, 2nd ed. McGraw-Hill Book Co., Toronto.

Winer, B. J. 1962. Statistical principles in experimental design. McGraw-Hill Book Co., Toronto.

SAS for OS/2

Serge Dupuis
Software Support
Alberta Public Works Supplies and Services
6950-113 Street
Edmonton, Alberta T6H 5T6

Abstract

The Statistical Analysis system (SAS) software has been developed to operate on several different computers. This includes Mainframes, Workstations and Personal Computers, each with its own operating system (TSO, VAX, DOS, UNIX and OS/2). This provides a great deal of flexibility. Users can run small analyses on DOS, access large databases on mainframes and move their SAS Programs to UNIX or OS/2 for large statistical or graphics applications. Each version of SAS has the same user language and procedures and is different only in the exploitation of each different operating system. This paper introduces some of these features.

OS/2 features

SAS for OS/2 is extremely flexible by exploiting features such as Named pipes, Dynamic data exchange (DDE) and others. Many of these are detailed in a recent paper in the IBM Personal Systems Developer (Goldstein, 1991)

Named Pipes allow SAS to establish many types of communication links. Within ONE machine, multiple SAS sessions can be running, with actions of one SAS session controlling another. With Named pipes, custom C applications can link to SAS. The support for Named Pipes allows for truly distributed processing in a networked environment. For example, SAS for OS/2 on one machine could be collecting real-time data from a serial data collection device, while another computer on the network could receive and analyse the data. Named pipes support is built into the SAS system and is independent of LAN software. It will run with any LAN software supporting named pipes, such as NOVELL for OS/2, IBM LAN server and Microsoft LAN manager.

Dynamic Data Exchange (DDE) is useful for establishing communications with other applications supporting DDE. Hot links can be established between EXCELL for example and SAS. As data is changed in the spreadsheet, SAS can analyse it directly.

Memory management

SAS is designed to take advantage of all available extended memory, up to 1 gigabyte of virtual memory. This memory is used for both executable code and data. The code generator used in the DATA step and the PROC code have been optimized for OS/2. Large statistical and graphics applications will show a marked improvement in performance with this increased memory.

Multitasking

OS/2 offers full multitasking, not time sliced tasking as used in DOS/WINDOWS. This means that any one application does not 'take over' the processor, as it does under WINDOWS. Also, there is full protection between applications, so that if one application fails, the entire computer does not need to be 'booted'.

Disk Access

OS/2 can use the High Performance File System (HPFS) which is optimized for multitasking and input/output throughput. SAS under OS/2 shows 25 to 50% increase in performance using HPFS, over the standard DOS File Allocation Table standard.

Comparison with Windows 3.0

SAS/PC under Microsoft Windows 3.0 will be available over the next year, however OS/2 and Unix are likely to be preferred by high-end users for complex applications. A paper comparing OS/2 and Windows 3.0 details some important differences in memory management, multitasking and communications (Gates, 1990). Under Windows, data and allocated memory are shared amongst applications which can lead to data corruption. Under OS/2, there can be up to two or three SAS sessions, while under Windows, only one SAS session is recommended. Limits to SAS datasets are the same for both Windows and OS/2, there is no limit to the number of observations but each observation has a maximum of 30,700 variables of 2 bytes each (or less variables of greater length).

References

Goldstein, P. (1991) 'Spotlight on SAS', IBM Personal Systems Developer, Spring 1991, 10-20

Gates M (1990), 'Comparing the SAS System under OS/2 and Microsoft Windows', SAS Internal documentation

Analysis of Air Monitoring Data

L.Z. Florence and H. Bertram
Alberta Environmental Centre
Postal Bag 4000
Vegreville, Alberta T0B 4L0

Abstract

Pollution monitoring has resulted in many measurement programs. Monitoring networks are used mainly to understand dispersion processes and measure deposition rates and/or levels. This paper is prepared as an introduction to the types of networks, specifically in Alberta, and the characteristics of spatial and temporal data which one might expect to analyze or summarize; examples come from dry sulphate and SO_2 deposition networks. An outline of methods and literature sources is presented for the statistical treatment of air monitoring data.

Introduction

Public concern for air quality and atmospheric deposition has resulted in numerous measurement programs, ranging from single station short-term research projects to long-term nation-wide monitoring programs. Current technology limits the scope of the monitoring programs to determining atmospheric concentrations; however, deposition models and deposition measurement procedures are rapidly developing.

The monitoring networks are used for at least two purposes: first, to help in our understanding of the process (emission, transport and deposition), and second, to facilitate a specific job such as to verify a predicted concentration gradient, or deposition rate. Short range monitoring is usually more successful because the pollutant does not have time to undergo chemical transformation or deposition to the earth's surface. Long range monitoring has to take into account surface deposition and both primary emissions and secondary emission products, such as the formation of secondary sulphate in the form of ammonium sulphate aerosol from sulphur dioxide emissions.

This paper presents discussion and examples of (1) types of monitoring networks, (2) characteristics of monitoring data, and (3) methods for analyzing monitoring data.

Types of Monitoring Networks

Monitoring networks tend to fall into three broad categories (Whelpdale 1983): exploratory, where they are alerting the public and scientific communities to the problem (ozone/nox/hydrocarbon and greenhouse gas networks), assessment, where methods are being refined and the extent of the problem is being defined (such as the acid deposition network) and regulatory, where the effectiveness of the regulations are being assessed (such as the urban lead network).

Air monitoring data for Alberta are collected from sites around the province with the intent of determining the effects of population growth and industrialization on the environment. Pollutants of concern are determined by their toxic/hazardous properties and the ratio of anthropogenic emissions to natural sources of emissions. Industries, and the provincial government as a quality control check, maintain monitoring networks in the vicinity of major processing plants as part of the regulatory and licensing requirements.

Network data are characterized by spatial distribution (location), time (date sampled), and intensity (atmospheric concentration). The type of sampling procedure must be taken into account in judging the appropriateness and reliability of the data. For example, continuous analyzers are used to determine violations of the one hour exposure limit, and monthly integrated samplers are used for assessing long-term emission trends and wind-rose type of spatial distributions.

Alberta Environment maintains networks in urban areas; for example, Edmonton has a twenty-seven site sulphation network and Fort McMurray has a six site network (Myrick and Asquin 1991). The Department also has sites which act as quality control checks around major processing plants.

Monitoring SO_2 Levels in a Network Using Passive, Integrated Samples

An established type of sampling device used in monitoring networks is the sulphation candle. The sulphation candle, which has been in use for the past fifty years, consists of a lead dioxide (peroxide) paste, using gum tragacanth as adhesive, painted in a cylindrical strip around the middle of a glass jar. The candle is exposed in the field for one month, where the lead dioxide reacts with ambient sulphur dioxide, hydrogen sulphide and mercaptans to form lead sulphate. After exposure the candles are returned to the laboratory for analysis.

A refinement of this device is the Huey plate, which is more convenient, less expensive and has some improvement in accuracy. Accuracy is improved by using a precise amount of lead dioxide for each sampler, thus eliminating some of the "blank" variability. Huey plates are prepared by pipetting a paste consisting of lead dioxide, gum tragacanth and cellulose fibres into a petri dish and then drying the plate in an oven. The sampler is then sent to the field, and exposed open side down in a convenient holder.

A more recent development in passive sulphur dioxide monitors is the use of potassium carbonate/glycerol on cellulose filter paper as the absorbent. As well, diffusion screens are now being used to eliminate wind velocity effects. Extensive networks of these low cost devices have been established in Alberta. The data gathered from these devices are used to delineate major areas of sulphur dioxide impact, and to identify trends in deposition levels.

Analyzing and Summarizing Monitoring Data: examples using sulphur dioxide (SO_2) emissions data from Alberta

Gilbert (1987), and references therein, discusses statistical methods for assessing spatial and temporal distributions of monitoring data. The reader is referred to these sources for more detailed discussion of the methods suggested here.

Spatial Patterns

Integrated samplers (total monthly deposition) in a network provide data that can be summarized monthly or yearly. Commonly applied methods for assessing spatial patterns are trend surface analysis and kriging, a weighting technique utilizing the correlation structure among locations in a grid or network. An example of how average annual dry sulphate deposition may be summarized using data from local networks represented by concentration isopleths is shown in Figure 1 (Fort McMurray Regional Air Quality Task Force, 1989) for the Suncor (N=40 sampling sites) and Syncrude (n=40 sampling sites). Note the concentration gradients with distance from each plant installation.

Spatial and Temporal Patterns

Figure 2 shows comparisons among four of the 27 SO_2 sampling locations in the Edmonton network. Each observation is an integrated, monthly concentration of SO_2 from January 1980 through December 1989. These plots, therefore, display both historical temporal variation and among location (within network) variability.

To model these relatively long data series, a Box-Jenkins (Box and Jenkins 1976) approach could be applied (requiring at least 50 data points), using software such as SAS/ETS (SAS Institute, 1988), for detecting trends or forecasting future deposition rates.

Referring to Figure 2, we might conclude that all four locations experienced a declining trend in SO_2, starting with the 1980 data continuing until at least the mid-1980's. The latter half of the 1980's data appear to fluctuate in a random, stationary manner. Gilbert (1987), chapters 16 and 17, provide a very useful guide for methodically assessing trends in monitoring data. Most of the techniques discussed by him are robust enough that outliers, missing and/or non-detectable data, are permissable.

A protocol for analyzing the data shown in Figure 2 might follow these steps:

1. First, plot the data; graphical analysis can be very powerful for not only assessing whether data contain patterns, but also as a quality control for detecting data errors or exceptional observations which should receive special attention.

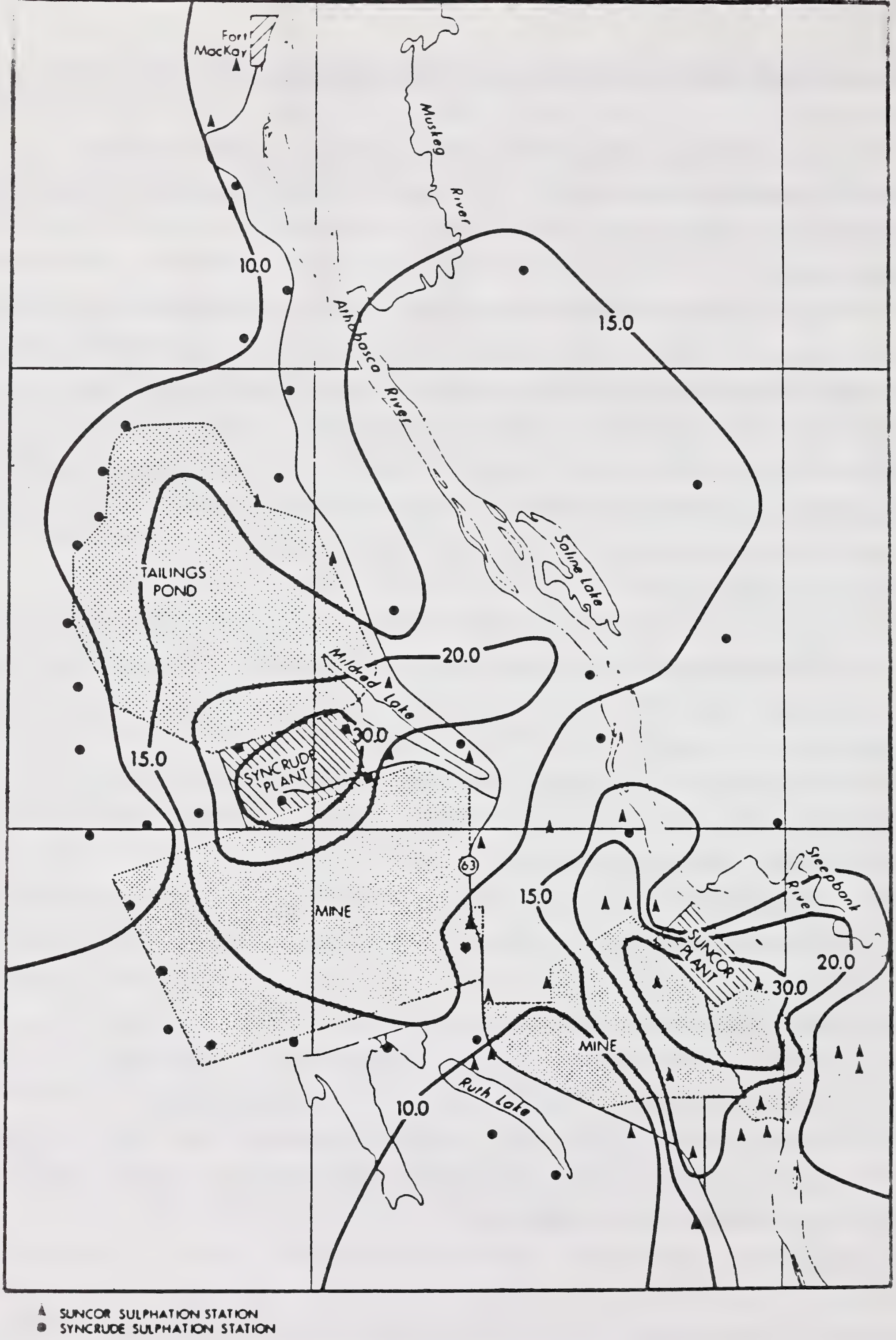


Figure 1. Estimates of average annual dry sulphate deposition (kg ha^{-1}) derived from information obtained from Suncor and Syncrude sulphates data collected from 1979 to 1983. (Figure 2, p. 10, Dabbs, 1985).

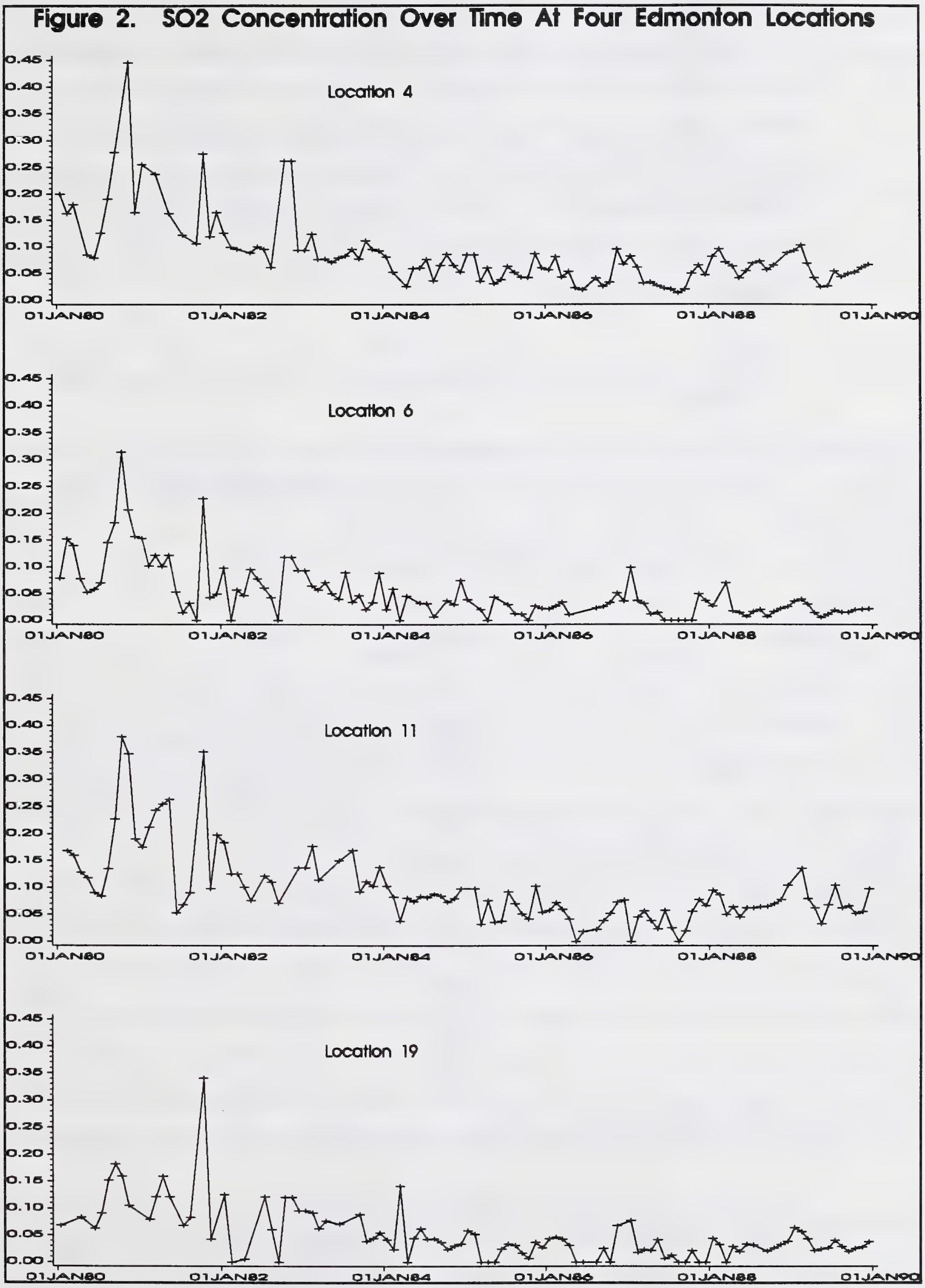


Figure 2. SO$_2$ Concentration Over Time at Four Edmonton Locations.

2. Once plotted, for example, in Figure 2, we may wish to test the null hypothesis that no trends exist over time and among locations. Total, global variation in SO_2 can be partitioned and the null hypothesis tested, as seen in Table 1, Part A. Because we would expect only a 0.021 chance of obtaining a larger chi-square value for "Station", we can ignore "Trend" and conclude that statistically significant different trend directions occurred between at least two of the four Edmonton stations during the 10 year period plotted in Figure 2. Table 1, Part B, indicates that all four locations experienced significant trends and, as might be expected from Figure 2, the weakest statistically was at #19.

Table 1. Tests for homogeneity of trends among four Edmonton locations; data are displayed in Figure 2.

Part A. Tests for homogeneity of trends among four Edmonton locations.

Source	Chi-square	df	$P>\chi^2$ [1]
Total	143.62	48	0.000
Homogeneity	55.58	47	0.183
Season	10.38	11	0.496
Station	9.75	3	0.021
Station-Season	35.45	33	0.354
Trend	88.04	1	0.000

Part B. Tests for trend at each Edmonton Location.

Location	Chi-square	df	$P>\chi^2$
4	25.96	1	0.000
6	40.94	1	0.000
11	26.28	1	0.000
19	4.60	1	0.032

[1] Probability of a larger chi-square value.

3. A further point of inquiry would likely be whether the trends remained significant upon adjusting for seasonal influence and, if so, what direction were the trends? Table 2, Part A, shows that seasonally adjusted trends existed and all were significant at the 0.03 level or less. Table 2, Part B, contains the Seasonal-Kendall slope coefficients and 95% confidence intervals for each location; as might be predicted from examining Figure 2, all four slopes are negative and the

weakest trend is at Station 19. We can conclude that the relative, median rate of change in SO_2 concentration at location #19 was approximately one third that observed at the other three stations.

Table 2. Tests for seasonally adjusted SO_2 trends at four Edmonton locations.

Part A. Tests for seasonally adjusted trends.

Location	Seasonal Kendall Score	P > SK[1]
4	-5.08	0.000
6	-6.38	0.000
11	-5.10	0.000
19	-2.13	0.033

Part B. Slopes and their confidence limits.

Location	Lower Limit[2]	SK Slope	Upper Limit
4	-0.011	-0.008	-0.005
6	-0.010	-0.008	-0.005
11	-0.013	-0.009	-0.006
19	-0.007	-0.003	0.000

[1] Probability of a greater Seasonal-Kendall (SK) score.

[2] Lower and Upper 95% confidence limits.

An obvious question arises as to the cause for the decline in SO_2 concentrations. Two obvious conclusions are: (i) industry was successfully reducing emissions, or (ii) due to the severe downward economic change in Edmonton about 1981-82, emissions declined due to industry shut-downs. Likely, both these reasons are part of the explanation, but a third is also very important because we also know that prior to 1982, a different lab was analyzing the SO_2 samples, and that it was using a different method of assay than was used after the change in labs. This type of information is invaluable when attempting to attribute cause-effect in trend analysis.

4. Seasonal trends are common in SO_2 monitoring data. For illustration, 10 means and medians, of 1980-89 monthly data, for Edmonton Location #4, are shown in Figure 3 (smoothing was done with a cubic spline algorithm available in SAS/STAT [SAS Institute, 1988]). We show both these measures of centrality to demonstrate that the median is less sensitive to extreme values than the mean (note especially August through November), thus better using all data rather than possibly

deleting obvious outliers. When the data are symmetrically distributed (note February, April and June) the mean and median are near the expected values for normally distributed populations. Figure 3 clearly shows that SO_2 concentrations increase in the fall-winter and are considerably reduced in the spring-summer. These seasonal trends are largely due to factors such as lower mixing heights, slower rate of oxidation to sulphate as hydroxyl radical concentrations decrease, and lower "washout" during the winter months, than during summer, resulting in greater ambient SO_2 deposition. Similar seasonality was observed in the other Edmonton stations (not shown).

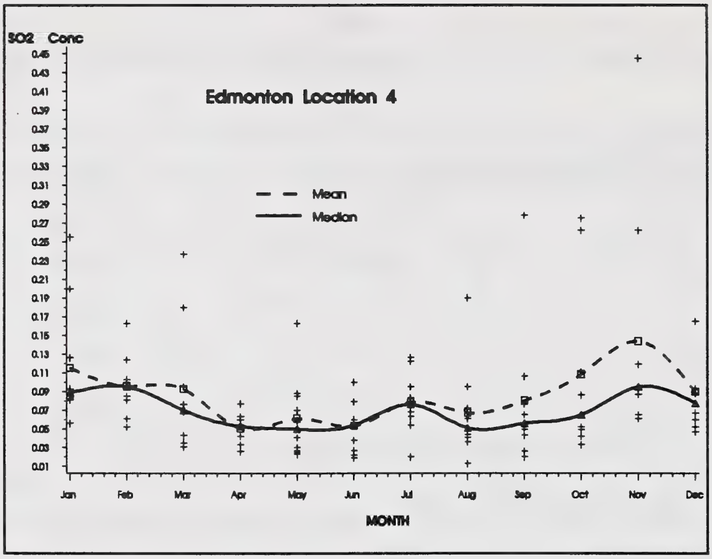


Figure 3. Variability in SO_2 concentration about the mean and median, over 10-year period, at one sampling location 4 in Edmonton, Alberta.

Summary

This paper has been an introduction to:

1. The types and purposes of air monitoring networks;
2. Examples, using SO_2 data from Alberta, showing how data may be distributed over time (years and seasons) and among locations;
3. And a stepwise approach, albeit not comprehensive, to analyzing monitoring data.

Acknowledgements

We are most appreciative for the computing assistance from John Kirtz, wordprocessing and manuscript preparation by Grovenore Flato and permission from Frank Ryan, Syncrude Canada, to reprint Figure 1.

References

Box, G.E.P. and G.M. Jenkins. 1976. Time series analysis: forecasting and control, 2nd ed. Holden-Day, San Francisco.

Dabbs, D.L. (editor). 1985. Atmospheric emissions monitoring and vegetation effects in the Athabasca oil sands region. Environmental Research Monograph 1985-5. Syncrude Canada Ltd.

Fort McMurray Regional Air Quality Task Force. 1989. Fort McMurray Regional Air Quality Assessment 1975-1986.

Gilbert, R.O. 1987. Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold Company, New York, 320 p.

Myrick, R.H., and D.H. Asquin. 1991. Air Quality Monitoring Report for Alberta, 1990. Air Assessment Section, Alberta Environment.

SAS Institute Inc. SAS/ETS User's Guide, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1988. 560 pp.

SAS Institute Inc. SAS/STAT User's Guide, Release 6.03 Edition, Cary, NC: SAS Institute Inc., 1988. pp.

Whelpdale, D.M. 1983. Monitoring and Assessment in Canada. Proceedings of Symposium on Monitoring and Assessment of Airborne Pollutants with Special Emphasis on Long Range Transport and Deposition of Acidic Materials. p. 25. NRCC Publication #20642.

Appendix A

Speakers List

1991 All Alberta Applied Statistics and Biometrics Workshop
Speakers List

Bertram, Henry	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Dupuis, Serge	PWSS Software Support Branch	2nd Floor, 6950-113 St. Edmonton, AB T6H 5V7
Florence, L. Zack	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Gilbert, Dr. Richard	Battelle, Pacific Northwest Laboratories	P.O. Box 999, Richland, WA 99352
Goonwardene, Laki	Alberta Ariculture Animal Industry Division	#204, 7000-113 St., Edmonton, AB T6H 5T6
Kozub, Gerry	Agriculture Canada, Research Station	P.O. Box 3000 Main, Lethbridge, AB T1J 4B1
Milliken, Dr. George	Kansas State University Department of Statistics	KSU, Dept of Statistics, Dickens Hall, Manhattan, KS 66506
Schaalje, Bruce	Agriculture Canada, Research Station	P.O. Box 3000 Main, Lethbridge, AB T1J 4B1
Schipper, Casey	Alberta Agriculture Health Management Branch	P.O. Box 4070, Station F, 6909-113 St., Edmonton, AB T6H 4P2
Weingardt, Ray	University of Alberta Dept. of Animal Science	3-10 Agriculture Forestry Bldg. Edmonton, AB T6A 2P5

Appendix B

Participants List

1991 All Alberta Applied Statistics and Biometrics Workshop
Participants List

Aku, Peter	University of Alberta Department of Zoology	CW-312 Bio Sciences Bldg., Edmonton, AB T6G 2E9
Asquin, Dave	Alberta Environment Environmental Quality Monitoring Branch	6th Floor, Oxbridge Place 9820 - 106 St., Edmonton, AB T5K 2J6
Bakowsky, Olenka	Alberta Forest Service	4th Floor, Bramalea Bldg. 9920-108 St Edmonton, AB T5K 2M4
Basarab, John	Alberta Agriculture Animal Industry Division	#204 J.G. O'Donoghue Bldg. 7000-113 Street Edmonton, AB T6H 5T6
Beck, Ron	Lethbridge Community College	3000 College Drive South Lethbridge, AB T1K 1L6
Bertram, Henry	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Booker, Clavin	Department of Herd Medicine Western College of Veterinary Medicine	University of Saskatoon, Saskatoon, SK S7N 0W0
Borchert, Trudy	Alberta Agriculture	#204, 7000-113 St., Edmonton, AB T6H 4T6
Brown, Richard	University of Alberta	#2,10038-110 St. Edmonton, AB T5K 1J6
Chalupa, David	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Conrad, Daniel	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Darroch, Barbara	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Das, Normal	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Deschamp, Kerry	Alberta Forest Service	9th Floor, Bramalea Bldg. 9920-108 St., Edmonton, AB T5K 2M4
Dew, Brenda	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Dieken, Fred	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Dinwoodie, Gordon	Alberta Environment Waste and Chemicals Division	5th Floor, Oxbridge Place 9820-106 St. Edmonton, AB T5K 2J6
Dombroski, Emil	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4

Dupuis, Serge	PWSS Software Support Branch	2nd Floor, 6950 - 113 St. Edmonton, AB T6H 5V7
Edmonds, Janet	Alberta Fish and Wildlife	Ste. 108, 111-54 St. Edson, AB T7E 1T2
Entz, Toby	Agriculture Canada	Box 3000 Main, Lethbridge, AB T1J 4B1
Esau, Rudy	Weed Scientist Alberta Special Crops and Horticulture, Alberta Agriculture	Brooks, AB T0J 0J0
Evans, Peter	Millar Western Pulp Ltd.	P.O. Box 1072, Whitecourt, AB T7S 1N9
Field, Kelly	University of Alberta	11305-68 St., Edmonton, AB T5B 1N8
Florence, L. Zack	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Gaudet, Irene	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Gilbert, Dr. Richard	Battelle, Pacific Northwest Laboratories	P.O. Box 999, Richland, WA 99352
Godby, Gavin	University of Alberta Dept. of Animal Science	3-10 Agriculture Forestry Blds. Edmonton, AB T6G 2P5
Goonwardene, Laki	Alberta Agriculture Animal Industry Division	#204, 7000 - 113 St., Edmonton, AB T6H 5T6
Goski, Bernie	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Gunter, Bill	Alberta Research Council Oil Sands and Hydrocarbon Recovery	P.O. Box 8330, Station F Edmonton, AB T6H 5X2
Hardin, Bob	University of Alberta Department of Animal Science	3-10 Agriculture Forestry Building Edmonton, AB T6G 2P5
Henry, Philip J.	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Hermesh, Reinhard	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Hiley, Jim	Agriculture Canada	6th Floor, Terrace Plaza Tower 4445 Calgary Trail South Calgary AB T6H 5A9
Hobson, David	Alberta Fish and Wildlife	Ste. 108, 111-54 St. Edson, AB T7E 1T2

Hop, Haakon	University of Alberta Department of Zoology	CW-313 BioSci Bldg., Edmonton, AB Edmonton, AB T6A 2E9
Hrynyk, Alan	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
James, Wendell	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Jenkins, Heather	Cross Cancer Institute Department of Epidemiology	11560 University Avenue Edmonton, AB T6G 1Z2
Johnson, C. Ian	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Jorgenson, Jon	Alberta Fish and Wildlife	#200, 5920-1A St. S.W. Calgary, AB T2H O63
Khan, A. Aziz	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Kirtz, John	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Kovacevich, Barbara	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Kozub, Gerry	Agriculture Canada Research Station	P.O. Box 3000 Main Lethbridge, AB T1J 4B1
Kryzanowski, Len	Alberta Agriculture Soils Branch	#905, 6909 - 116 St. Edmonton, AB T6H 4P2
Kumar, Yogesh	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Lakusta, Tom	Alberta Forest Service Timber Management Branch	9920 - 108 St., Edmonton, AB T5K 2M4
Lam, Tonya	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Lamontagnf, Sebastian	University of Alberta	CW-313 BioSci Bldg, Edmonton, AB T6A 2E9
Lennon, Joe	Planning Division Alberta Environment	9th Floor, Oxbridge Place 9820-106 St Edmonton, AB T5K 2J6
Lewis, Barry J.	W.R. Dempster & Associates	#201, 104-176 St., Edmonton, AB T5S 1L3
Liu, M.F.	University of Alberta Dept. of Animal Science	3-10 Agriculture Forestry Bldg. Edmonton, AB T6G 2P5
Lucyk, Dale	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Lyka, Wendy	Millar Western Pulp	Box 1072, Whitecourt, AB T5S 1N9
Mah, Zeva	Alberta Cancer Board	#120 1040 - 7th Ave. S.W. Calgary, AB T2P 3G9

McKenzie, Colin	Alberta Agriculture	ASCHRC, Brook, AB T1R 1E6
Milliken, Dr. George	Kansas State University Department of Statistics	KSU, Dickens Hall Manhattan, KS 66506
Morley, Paul	Dept. of Veterinary Internal Med., Western College of Veterinary Medicine	University of Saskatchewan Saskatoon, SK S7N 0W0
Muhlenfeld, Angela	Alberta Bureau of Statistics	Suite 600, 10611-98 Ave. Edmonton, AB T5K 2R7
Myrick, Bob	Alberta Environment Environmental Quality Monitoring Branch	6th Floor, Oxbridge Place 9820 - 106 St, Edmonton, AB T5K 2J6
Nguyen, Chung	University of Alberta	442 Earth Science Bldg. Edmonton, AB T6G 2E3
Naazie, Augustine	University of Alberta Dept. of Animal Science	3-10 Agriculture Building Edmonton, AB T6G 2P5
Nietfeld, Marie	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Parraquez, Carlos	Alberta Environment	9th Floor, Oxbridge Place 9820-106 St., Edmonton, AB T5K 2J6
Paul, Andrew	University of Alberta Dept. of Zoology	Edmonton, AB
Phillips, Paul	W.R. Dempster & Associates	#201, 10464-176 Street Edmonton, AB T5S 1L3
Precht, Dan	Senior Analyst Statistical Applications University Computing Systems	231 General Services Building University of Alberta Edmonton, AB T6G 2H1
Prior, Michael	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Schaalje, Bruce	Agriculture Canada Research Station	P.O. Box 3000 Main Lethbridge, AB T1J 4B1
Schilf, Janet	Alberta Forest Service	4th Floor, Bramalea Bldg. 9920-108 St., Edmonton, AB T5K 2M4
Schipper, Casey	Alberta Agriculture Health Management Branch	P.O. Box 4070, Station F 6909-113 St., Edmonton, AB T6H 4P2
Schreiner, Kurt	Alberta Bureau of Statistics	600 Park Plaza, 10611-98 Ave. Edmonton, AB T5K 2R7

Selirio, Sid	Alberta Hail and Crop Insurance Corporation	Bag Service #16, Lacombe, AB T0C 1S0
Serink, Brian	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Sharma, Paul	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Sherstabetoff, Rick	Alberta Agriculture Soils Branch	#205, J.G. O'Donoghue Bldg. 7000-113 St. Edmonton, AB T6H 5T6
Skinner, Frank	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Smith, Kirby	Alberta Fish and Wildlife	Ste. 108, 111-54 St. Edson, AB T7E 1T2
Somers, Jim	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Taerum, Terry	University of Alberta	U.S.C., Edmonton, AB T6G 2E1
Thomson, David G.	Alberta Research Council Environmental Research and Engineering	6815 - 8 St., N.E., Calgary, AB T5K 7H7
Thorlakson, B.	Animal Research International	P.O. Box 3490, Airdrie, AB T4B 2B7
Tong, Alan	Agriculture Canada	Bag Service 5000, Lacombe, AB T0C 1S0
Urso, Alessandro	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
VanDonkersgoed, Joyce	WDO	124 Veterinary Road, Saskatoon, Sk S7K 0W0
Webb, Debbie	University of Alberta	CW-313 BioSci Bldg., Edmonton, AB T6A 2E9
Weingardt, Ray	University of Alberta Dept. of Animal Science	3-10 Agriculture Forestry Bldg. Edmonton, AB T6A 2P5
Wong, Roy	Alberta Research Council	250 Karl Clark Road, Edmonton, AB T6H 5X2
Wu, Shaole	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Xu, J. G.	Department of Soil Science University of Alberta	Edmonton, AB T6G 2E3
Yee, Dennis	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Yeung, Paul	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4

Sillito, Ed	Alberta Hail and Crop Insurance Corporation	Bag Service #10, Lacombe, AB T0C 1S0
Sorrey, Brian	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Sharma, Paul	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Shwalukoff, Rick	Alberta Agriculture Soils Branch	#905, J.G. O'Donoghue Bldg. 7000-113 St. Edmonton, AB T6H 5T6
Skinner, Frank	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Smith, Kirby	Alberta Fish and Wildlife	Stn. 600, 111-54 St. Edson AB T7E 1T2
Stewart, Jan	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Thomas, Terry	University of Alberta	U.S.C., Edmonton, AB T6G 2E1
Trottier, David G.	Alberta Research Council Environmental Research and Engineering	6815 - 8 St. N.E., Calgary, AB T2E 7H7
Thompson, B	Antibody Research International	P.O. Box 3990, Airdrie, AB T4B 2B3
Tong, Alan	Agriculture Canada	Bag Service 5000, Lacombe, AB T0C 1S0
Urso, Alessandra	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
VanDonkersgoed, Joyce	VIDO	120 Veterinary Road, Saskatoon, SK S7N 0W0
Webb, Dennis	University of Alberta	CW 315 Biology Bldg., Edmonton, AB T6G 2E9
Weingardt, Ray	University of Alberta Dept. of Animal Science	4-10 Agriculture Forestry Bldg. Edmonton, AB T6G 2P5
Wong, Ray	Alberta Research Council	250 Karl Clark Road, Edmonton, AB T6H 5X2
Wu, Shiang	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Xu, C.C.	Department of Soil Science University of Alberta	Edmonton, AB T6G 2E3
Yu, Dennis	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4
Young, Paul	Alberta Environmental Centre	Bag 4000, Vegreville, AB T9C 1T4